Summer 2018

Isolation, sequencing, and characterization of four transmissible antibiotic resistance plasmids captured from bacteria in stream sediments

Curtis J. Kapsak

Isolation, Sequencing, and Characterization of Four Transmissible Antibiotic Resistance Plasmids Captured from Bacteria in Stream Sediments

Curtis J. Kapsak

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Science

Department of Biology

August 2018

FACULTY COMMITTEE:

Committee Chair: Dr. James B. Herrick

Committee Members/ Readers:
Dr. Morgan Steffen
Dr. Timothy Bloss
Dr. Stephen Turner

Acknowledgments

The completion of thesis would not have been possible without the support from Dr. James Herrick, my peers, my friends, and my family. Dr. Herrick has been an amazing mentor to work under for the past 5 years and I have learned so much from him. I will always treasure my time at JMU working in his laboratory. This also would not have been possible without the support from various others – Dr. Stephen Turner and Dr. Brian Capaldo for their bioinformatics support, Dr. Steve Cresawn, for setting me up to use his server. I'm grateful to Erika Gehr for capturing pEG1-06 and for her help with the exogenous plasmid capture. I'm also grateful for Kevin Libuit for laying the groundwork with plasmid sequencing in general and for his advice and help throughout the completion of this project.

Table of Contents

List of Tables

List of Figures

Chapter 1 Figures

Chapter 2 Figures

Abstract

Self-transmissible plasmids are key vectors in the transfer of resistance, catabolic, and other genes among bacteria native to environments such as streams and wetlands. The evolution of antibiotic resistance in particular is known to be powerfully affected by conjugative plasmid transfer due to the ease in which some plasmids can be horizontally transferred into a broad range of host bacteria and their ability to exchange mobile genetic elements that often contain antibiotic resistance genes.

In this study, we captured tetracycline resistance plasmids from stream sediments impacted by agricultural runoff. We selected for resistance plasmids using tetracycline, an antibiotic commonly used in agricultural operations, due to the numerous neighboring cattle pastures and poultry farms. We hypothesized that stream sediment is a "hot spot" for horizontal gene transfer due to the use of antibiotics in agricultural operations combined with runoff into streams. Selective pressures exerted on gut and fecal bacteria of farm animals may select for antibiotic resistance genes that can be horizontally transferred to native stream sediment bacteria when runoff events occur.

We characterized four transmissible, tetracycline resistance plasmids: the 71 kb IncP-1β plasmid pEG1-06, the 121 kb IncA/C_2 plasmid pCCRT11-6, and the 59 kb IncP-9 plasmids pCCP1 and pCCP2. We built upon and improved the methods developed for the preparation of plasmid DNA for sequencing using 2nd and 3rd generation DNA sequencers, hybrid genome assembly, annotation, and analysis. We demonstrated this process by assembling the four plasmid genomes into single, circular contiguous sequences and compared them to the closest related plasmids allowing us to classify their

respective incompatibility groups, reveal the essential backbone and accessory genes present on the plasmid genomes including antibiotic resistance genes, and determine their similarity to the closest related known, existing plasmids.

General Introduction

Plasmids. Plasmids are circular elements of double-stranded DNA found in bacteria, archaea, and some eukaryotes that are normally replicated independently of chromosomal DNA. Many – those known as *self-transmissible* or as *mobilizable* plasmids – can be horizontally transferred between mature bacteria by conjugation. Genes commonly found in plasmids often encode traits that help a bacterium adapt and survive in its environment, such as antibiotic resistance genes, virulence genes, and other unique genes. Due to the mosaic nature of plasmids they can also serve as vehicles for other mobile and mobilizable genetic elements such as integrons and transposons, which adds yet another level of potential recombination and mobility.

A self-transmissible or conjugative plasmid is defined as a plasmid that contains genes for all four components of a membrane mating pair formation complex: (1) an origin of replication, (2) a relaxase, (3) a type IV coupling protein, and (4) a type IV secretion system. A mobilizable plasmid is defined as a one that only contains genes for three of four components of a conjugative complex and lacks the genes for a type IV secretion mating channel which is required for the conjugation process. Mobilizable plasmids differ from self-transmissible plasmids in that they rely upon the presence of a mating channel from another genetic element (e.g. a self-transmissible plasmid) present in the cell in order for horizontal gene transfer (HGT) to occur (Smillie *et al.*, 2010).

Plasmids are typically classified either by their incompatibility group or by mobility typing (MOB). Plasmid incompatibility is defined as the failure of two co-resident plasmids to be stably inherited together in the absence of external selection

(Novick *et al.*, 1976). If two plasmids exist in the same cell and are of the same incompatibility group, only one will be replicated and vertically passed down to its progeny. Plasmid typing based on incompatibility groups has been the gold standard for plasmid classification for many years; however due to the mosaic nature of plasmid genomes (i.e. plasmids are dynamic and plastic in their ability to gain or lose genes through horizontal gene transfer), the classification of plasmids using incompatibility has not always reflected their true evolutionary relationships (Alvarado *et al.* 2012). Highly similar plasmids can be compatible while largely non-homologous plasmids can be incompatible (Alvarado *et al.*, 2012). A newer method to classify plasmids is based on Degenerate Primer Mobility Typing (DPMT) (Alvarado *et al*., 2012; Garcillan-Barcia *et al.*, 2015). DPMT is a PCR-based typing strategy that aims to characterize plasmids based on evolutionary relationships among relaxase genes (MOB genes), because these are the only genes that are common to all transmissible and mobilizable plasmids (Garcillin-Barcia *et al.*, 2015).

IncP plasmids. Plasmids belonging to the IncP incompatibility group are a well-characterized group of plasmids due to their high prevalence, their clinical relevance due to their often containing multiple antibiotic resistance genes, and their broad host range. IncP plasmids are described as "promiscuous" plasmids due to their ability to replicate and be stably maintained in almost all Gram-negative bacteria. They are often found in the Enterobacteriaceae in genera such as *Salmonella*, *Escherichia*, and *Klebsiella*, as well as *Pseudomonas* (a non-Enterobacteriaceae in which they are designated IncP-1) (Adamczyk and Jagura-Burdzy, 2003). IncP plasmids often contain genes that encode catabolic functions such as the degradation and utilization of chemical compounds such

as xylene or toluene (Dennis, 2005). They frequently contain a variety of antibiotic resistance genes conferring resistance to (for example) tetracyclines, sulfonamides, extended spectrum β-lactams, and aminoglycosides (Rozwandowicz *et al.*, 2018).

IncA/C plasmids. IncA/C plasmids are a group of low-copy number, self-transmissible plasmids that range in size from 40-230 kb, although smaller conjugative variants with sizes 18-25 kb have been reported (Rozwandowicz *et al.*, 2018). They have a broad host range and are typically found in members of the Beta-, Gamma-, and Deltaproteobacteria, including species such as *Escherichia coli, Klebsiella pneumoniae, Salmonella enterica, Vibrio cholerae,* and *Aeromonas hydrophila* (Harmer and Hall, 2015). These plasmids often carry genes of clinical significance such as antibiotic resistance genes conferring resistance to carbapenems, third-generation cephalosporins, sulfonamides, tetracyclines, and extended-spectrum beta-lactam antibiotics which can make treatement of a drug-resistant infection difficult if the causative agent has an IncA/C plasmid (Carattoli *et al.*, 2012). These resistance genes are often located on antibiotic resistance islands (ARI-A or ARI-B) which are associated with the global dissemination of extended-spectrum β-lactamase genes via IncA/C (Rozwandowicz *et al.*, 2018). The IncA/C plasmids are divided into two variants, A/C_1, which is typified by the IncA reference plasmid pRA1, and A/C_2 a variant that differs by 26 single nucleotide polymorphisms in the *repA* gene (plasmid replication protein A) (Rozwandoicz *et al.*, 2018). $IncA/C_2$ plasmids are further divided into two groups, type 1 and type 2, and these differ based on the amino acid length of the *rhs* gene (as a result of the accumulation of insertions and deletions) as well as the presence of ARI-A and ARI-B (ARI-A is only found in type 1, ARI-B found in both types) (Harmer and Hall, 2015).

Antibiotic resistance. Humans are currently participating in an antibiotic and antibiotic resistance arms race with bacteria, where bacteria develop resistance to the antibiotics used by humans to rid them of pathogenic bacteria causing disease. The rate of bacterial evolution is clearly outpacing the rate that humans are able to discover new and effective antibiotics. The highly mutable nature of bacterial genomes is a major contributor to the evolution of bacteria, allowing them to adapt to their environment and persist under unfavorable conditions (Davies and Davies, 2010).

The observed increase in antibiotic resistance is likely not only due to use, the misuse, and overuse of antibiotics as a treatment for human infection, but also to the overuse of antibiotics in agriculture (Wegener, 2003). The US Food and Drug Administration (FDA) Center for Veterinary Medicine reported that from 2009 to 2016 domestic sales and distribution of antimicrobials approved for use in food-producing animals increased by 11% (Center for Vet Medicine at FDA, 2017). In 2016, 13,983,016 kg (15,413.6 tons) of antimicrobials were sold while 42% of those sold were tetracyclines (Center for Vet Medicine at FDA, 2017). Sixty-nine percent of the antimicrobials sold in the US were approved by the FDA and labeled for use in both therapeutic (treatment of a bacterial infection) and non-therapeutic or production applications (increased weight gain) while 31% were labeled solely for therapeutic applications (Center for Vet Medicine at FDA, 2017).

The use of antibiotics and antimicrobials in animal husbandry over the past 30 years has actively selected for bacteria that possess genes conferring antibiotic resistance (Cantas *et al.*, 2013). These resistant populations of bacteria are present in feces and eventually may enter adjacent water sources via runoff from neighboring agricultural

areas. Bacteria introduced into a system via runoff are able to exchange mobile genetic elements with native bacteria using horizontal gene transfer (HGT) mechanisms (Fry and Day, 1990). Streams and stream sediment harbor large, diverse populations of native bacteria. If in close proximity to a farm, these populations of native bacteria that live in the stream water column or sediment are subject to periodic or continuous contact with fecal bacteria via direct deposition or runoff from applied fertilizer (Herrick *et al.*, 2014). Such contact could lead to the transmission of antibiotic resistance genes between non-pathogenic native bacteria and pathogenic (and opportunistic pathogenic) fecal bacteria.

In Virginia, the Department of Environmental Quality (DEQ) is tasked with monitoring and reporting the quality of the various bodies of water within the state. They determine that a waterbody is "impaired" when it contains more of a pollutant than is allowed by water quality standards. For all of the rivers and streams in the state that were assessed by the DEQ, 70% were designated as "impaired" due to high levels of fecal bacteria. In the Potomac-Shenandoah river basin, where this study was conducted, 79% of the rivers designated as "impaired" were due to the presence of fecal bacteria (DEQ, 2016).

Horizontal gene transfer. HGT is perhaps the primary cause of the rapid spread of antibiotic resistance (Revilla *et al.*, 2008). Bacteria use HGT, the transfer of DNA between mature cells, and vertical transfer of DNA to progeny, to adapt to stressful conditions and thrive in new environments (Smillie *et al.*, 2010). HGT can occur between both closely and distantly related species alike and can potentially occur between non-pathogenic and pathogenic bacteria (Shoemaker *et al.*, 2001). Exchange of antibiotic resistance genes from a non-pathogenic species to pathogenic species is one of the most

concerning potential outcomes of HGT. There are three classical mechanisms of horizontal gene transfer: transduction, transformation, and conjugation.

Transduction is the transfer of new genetic material into a bacterium by a bacteriophage (a virus that infects bacteria) that has replicated its genome within a donor bacterium. It can either transfer packaged, random DNA fragments made using the host's replication machinery or the DNA adjacent to the phage attachment site (Ochman *et al.*, 2000). Fairly large amounts of DNA can be transferred in a single event and is only limited by the size of the phage capsid (Ochman *et al.*, 2000). This form of HGT is less likely to occur than transformation or conjugation due to the fact that phages have a limited host range largely dependent upon the surface proteins found on their host that aid in attachment (Smillie *et al.*, 2010). Transduction is unique in that it does not require physical contact between the donor cell and the recipient cell. One of the more well-known examples of transduction is the acquisition of the genes that code for Shiga toxin production in pathogenic *Escherichia coli* O157:H7. This strain received these genes via a bacteriophage early in its evolutionary history and this transduction event caused a major change in its genome and its ability to cause disease in humans (Wick *et al.*, 2005).

The second mechanism of HGT is **transformation**, which is the uptake and incorporation of naked DNA from the environment into bacterial cells. Both chromosomal and extrachromosomal (plasmid) DNA can be transformed into bacteria and does not require a living donor. Most naked DNA is a result of cell lysis and can originate in cells that are not closely related or proximal to the recipient (Lorenz and Wackernagel, 1994). Not all bacteria are *competent*, or able to take up foreign DNA. Competency can either occur naturally or can be induced by a variety of artificial

treatments, such as heat-shock, $CaCl_2$, and electro-shock or electroporation (Chen and Dubnau, 2004). Induced competency increases cellular membrane fluidity by creating small pores in the membrane to allow for DNA (typically plasmids) to traverse the membrane.

The third route for HGT is **conjugation**. Physical contact between two bacterial cells is necessary to initiate gene transfer between donor and recipient cells (Gotz and Smalla, 1997). Typically, the form of DNA being exchanged during conjugation is a self-transmissible or mobilizable plasmid. Successful conjugation depends on the genetic information and is encoded by the plasmid itself. There are several genes involved in the conjugation mechanism which function to form the conjugative bridge - the pilus - between the donor and recipient. This mechanism is required to transfer and copy a plasmid into the recipient but may also be a conjugative transposon (Ochman *et al*., 2000).

HGT and antibiotic resistance. The ability of bacteria to transfer and rearrange genetic content to gain new traits has been clearly demonstrated in terms of antibiotic resistance (Smillie *et al.*, 2010). One recent instance of horizontal gene transfer of antibiotic resistance genes is the spread of colistin resistance among bacteria such as *E. coli* and *Klebsiella pneumoniae* (Liu *et al.*, 2015). Colistin belongs to the family of polymyxin antibiotics and has a broad spectrum of activity against many Gram-negative bacteria including most species of *Enterobacteriaceae* (Li *et al*., 2006). Doctors rely on colistin as a last resort antibiotic to treat carbapenem-resistant bacterial infections. Colistin has been used in sub-therapeutic amounts for animal growth promotion in some agricultural operations in China (Doyle *et al.*, 2013).

A recently discovered plasmid-mediated gene that confers resistance to colistin, *mcr-1,* was found in high abundance among *E. coli* isolates sampled from raw meat, animals, and infected patients in central and eastern China (Liu *et al.,* 2015). The authors observed *mcr-1* carriage in 15% of raw meat samples, 21% of animals, and 1% of patients across 13 different provinces in China. The most alarming aspect of this study was the apparent ease in which a plasmid containing the *mcr-1* gene was transferred among differing pathogenic bacterial species such as *E. coli, Pseudomonas aeruginosa,* and *K. pnemoniae*. The plasmid, pHNSHP45, can not only be transferred at a high rate, it has also been shown to be stable in these bacteria, even without the selective pressure of antibiotics (Liu *et al.,* 2015). This is harrowing news, as plasmid-borne colistin-resistance may inevitably spread among pathogenic bacteria which could potentially infect human populations.

Second generation DNA sequencing. The 2005 release of 454 Life Sciences' DNA sequencer, the 454 GS20, marked the advent of 2nd generation DNA sequencing (Marguilies *et al.,* 2005). The GS20 was the first commercially-available device that had the capability of massively parallel DNA sequencing. The device delivered up to 25 megabases (1 megabase = 1 million bases) per sequencing run (Margulies *et al.,* 2005), which was much higher DNA sequence throughput than first generation Sanger sequencing technologies.

Second generation sequencers differ from first generation sequencers in that they allow for the multiplexing of sequencing reactions. This means that the device can sequence multiple strands of DNA in parallel, thus allowing second generation sequencers to have much greater scalability (Shendure *et al.,* 2017). This key distinction

drove the cost of sequencing down immensely, as data could be generated much faster for a larger number of samples. In 2004, the cost of sequencing 1 megabase was approximately $1000 and in 2008, when many sequencing centers adopted and began using 2nd gen sequencers, the cost dropped to under $10 per megabase. In 2017, due to further advancements in sequencing technologies, the cost has plummeted even further to $0.012 per megabase (Wetterstrand, 2017).

The 454 Life Sciences lineup of sequencers and other 2nd generation sequencers from companies such as Solexa/Illumina and Ion Torrent, all operate using "sequencing by synthesis" approaches (SBS) (Morey *et al.,* 2013). First, DNA is clonally amplified by polymerase chain reaction (PCR), meaning that a DNA fragment is amplified into many identical copies which are then denatured into single-stranded DNA fragments. Second, the single-stranded, amplified fragments are spatially separated and immobilized on a surface, usually on a flowcell (Illumina) or a microchip (Ion Torrent) (Shendure *et al.*, 2017). Finally, single stranded DNA templates are subject to polymerization of their complementary strands, one nucleotide at a time, to allow for the detection of each nucleotide. Detection methods differ between the various technologies, but typically when a nucleotide is added to the complementary strand, a byproduct is released, such as hydrogen ions (H^+) or light (Morey *et al.*, 2013).

Illumina sequencing by synthesis. Illumina is aptly named for the light-based sequencing methods that are employed in their sequencers. During a sequencing run, fluorescently-labelled "reversible-terminator" deoxynucleotides (dNTP's) are used for polymerization of the complementary DNA strand (Turcatti *et al.,* 2008). Between each sequential nucleotide addition, lasers are used to excite the fluorophore that is attached to

the dNTP, and the signal is recorded using a camera. This signal is used to derive the nucleotide sequence that was incorporated into each individual DNA strand (Shendure *et al.*, 2017).

Second generation sequencers rely upon PCR for various processes throughout their respective sequencing workflows, and while PCR enables the generation of massive amounts of sequence data, it has its drawbacks. Clonal amplification of DNA can lead to amplification biases resulting in uneven uniformity of coverage and potential sequencing artifacts due to polymerase errors (Acinas *et al.*, 2005). To minimize the effects of these biases, clonal amplification reactions are performed with a reduced number of cycles (usually under 15 cycles) to limit polymerase errors (Asan *et al.*, 2011). The amplification of DNA also results in a loss of information in that post-transcriptional modifications (e.g. methylation) are lost post-amplification. The loss of nucleotide modifications is a hindrance if the investigator is specifically interested in modifications; however it is by and large an advantage due to the increase in basecalling accuracy. The increase in basecalling accuracy is due to the sequencer not having to differentiate between 5-mC and non-methylated cytosine signals.

Despite the low cost per base, high accuracy, and high throughput of second generation sequencers, the main disadvantage of these sequencers for the purposes of genome assembly, is the short read length of the data produced, which range anywhere from 35 - 500 bp (typically on the shorter side around 75-150 bp). Short reads, regardless of how high their accuracy is, can make genome assembly a challenge, especially when assembling a genome *de novo*. Genomes that contain long repetitive elements prevent the resolution of a genome and are especially difficult to assemble when the repetitive

element is longer than the length of the sequence reads and in the case of bacterial genomes, 2nd generation sequencers oftentimes cannot provide complete genome sequences (Pallen and Loman, 2015).

Third generation DNA sequencing. Third generation sequencers are distinct from 2nd generation sequencers in that they sequence native DNA molecules directly and not from a synthesized template. Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) Single Molecule, Real-Time (SMRT) sequencing technologies dominate the 3rd generation sequencing market as they produce sequence reads directly and at lengths much higher than 2nd generation sequencers. PacBio and ONT users report reads typically over 10 kb and up to 100 kb (Shendure *et al.*, 2017). In theory, nanopore sequencers have no limits to their read lengths and through the use of careful, high molecular weight DNA extractions and library preparation protocols users have obtained single "ultra-long" reads up 882 kb in length (Jain *et al.*, 2018). With increasingly longer sequence read lengths becoming more readily available and accessible, scientists are able to apply them to the long-standing problems associated with short read sequence data. Bacterial genomes have been assembled using long reads such as the 4.6 Mb *E. coli* K-12 MG1655 genome (Loman *et al.*, 2015), the 1 Mb *Rickettsia typhi* genome (Elliot *et al.*, 2018), and the 5.7 Mb *Klebsiella pneumoniae* genome (Wick *et al.*, 2017). Additionally, 3rd generation sequencers have allowed for the assembly of large and complex eukaryotic genomes such as *Arabadopsis thaliana* (135 Mb genome) and human (3.1 Gb genome) and result in assemblies that are much more contiguous and structurally accurate than those possible with short read data alone (Michael *et al.*, 2018; Jain *et al.* 2018).

Nanopore sequencing. Nanopore sequencers measure picoampere (pA) level changes in ionic current as single stranded DNA (ssDNA) molecules pass through a biological nanopore. Within the flowcells of all ONT sequencers are arrays of biological (i.e. non-synthetic) nanopores. These pores have a 1-2 μm opening that allows ssDNA molecules to pass through at a speed of 450 bases per second, and whose speed is guided by a motor protein that rests on top of the nanopore (Oxford Nanopore, 2016). The speed at which bases are sequenced combined with the multiplexing of sequencing across an array of 2048 nanopores controlled in groups of 512 channels allows for the routine generation of 3-5 Gb (some users report up to 18 Gb) of data per flowcell (Leggett and Clark, 2017).

One of the main advantages of nanopore sequencing is that does not require PCR amplification prior to or during sequencing, thus eliminating a potential source of bias in the sequence data. Another main advantage to this technology is that since it allows for the detection of nucleotide sequence directly from native molecules, it can also detect and discriminate base-modifications, such as methylated residues (Jain *et al.*, 2016) and for the first time in history, allows for the sequencing of RNA molecules directly without the requirement for conversion to cDNA (Garalde *et al.*, 2018).

Nanopore sequencing does have disadvantages, some of which will be discussed in detail in the first chapter. One of the main drawbacks is that since ONT sequencers do not require PCR prior to sequencing (although ONT does have library preparation kits for low-quantity samples that involve amplification), the required minimum quantity of starting DNA is high compared to 2nd generation sequencers, ranging from 400 ng to 1 μg depending on the library preparation kit that is used (compare to the range of 1 ng to 500

ng required for Illumina sequencing). Another disadvantage to ONT sequencing is that because the sequencer is essentially threading DNA through a biological protein, the nanopores are sensitive to contaminants in DNA samples and various other factors such as exposure to air or dramatic temperature changes. Therefore, ONT sequencers require that DNA samples are extremely pure, non-degraded, and free of contaminants such as phenol, ethanol, salts, EDTA, etc. that may degrade or denature the nanopores, which leads to lower quality data and lower throughput.

The biggest drawback to ONT sequencing is the accuracy of the reads, which is currently estimated to be ~86% for the median read identity (weighted by read length) with an average distribution ranging from 65-95% read identity for 1-dimensional (1D) reads (Wick *et al.*, 2018). This accuracy rate deters many researchers from using the technology, especially if their project requires a high accuracy rate of sequence data, but the data is useable for many other purposes, such as real-time detection of pathogenic bacteria or viruses (e.g. Ebola virus), hybrid short and long read *de novo* genome assembly, and detection of structural variants in cancer cell lines (Quick *et al.*, 2016; Wick *et al.*, 2017; Norris *et al.*, 2016).

Sequencing on Illumina and ONT platforms for hybrid assembly. While data from either a 2nd or a 3rd generation sequencer can be sufficient for generating a *de novo* draft genome assembly, both approaches can be used in conjunction to produce genome assemblies that are highly accurate in structure and sequence. Such a “hybrid” approach takes advantage of the high accuracy rate of the short reads and the read lengths of the long reads. There are two main approaches to a hybrid assembly, the first of which is to assemble using the long reads first and perform error correction using the short reads.

Tools such as Miniasm (an overlap layout consensus-based assembler) and Pilon (which polishes draft genome assemblies by using high accuracy short reads) have been developed for this purpose (Li, 2016; Walker *et al.*, 2014). The second approach is to assemble the short reads first and scaffold the contigs using the long reads. Tools such as Unicycler have been developed for this purpose, specifically for *de novo* assembly of bacterial genomes and plasmids (Wick *et al.*, 2017).

In this study, we build upon and improve the methods developed by Libuit (2016) for the preparation of plasmid DNA for sequencing, plasmid genome assembly, annotation and analysis of large, antibiotic resistance plasmids using 2nd and 3rd generation DNA sequencers. We demonstrate this process by sequencing four tetracycline resistance-encoding plasmids that vary in size, incompatibility group, and copy-number that were captured without cultivation of the host(s) from streams located in the Shenandoah Valley. The genomes of self-transmissible tet^{R} plasmids pEG1-06, pCCRT11-6, pCCP1, and pCCP2 were sequenced, assembled *de novo*, annotated, and compared to the closest related plasmids allowing us to classify their respective incompatibility groups, reveal the various genes present on the plasmids, and determine their similarity to the closest related plasmids.

Chapter 1

A method for preparing isolated plasmid DNA for sequencing on Illumina and Oxford Nanopore Technologies sequencers

Introduction

Whole genome sequence (WGS) data of bacterial isolates contain not only chromosomal sequence content but will additionally contain plasmid sequence if the isolate harbors one or more plasmids that are sequenced alongside the chromosomal DNA. Resolving plasmid genomes from WGS data poses many challenges, because plasmids commonly share repetitive sequences with the chromosome, making plasmid genome assembly difficult, especially when short sequence reads are used (Arredondo-Alonso *et al.*, 2017). Long reads that are capable of spanning beyond these shared repetitive regions can be assembled in a "hybrid" manner with the short reads to solve these assembly issues (Wick *et al.*, 2017); however there is the possibility that in the process of preparing libraries and sequencing them, plasmids can be missed or excluded altogether and this may occur for a number of reasons. One could be due to sheer probability, in the case that a plasmid is low-copy number and each cell contains one plasmid copy per cell. The sheer size and abundance of the chromosome may outweigh a small plasmid and it could be missed. Another possibility is that a plasmid's supercoiling states may prevent it from being incorporated into sequencing libraries, due to no free ends being available for the ligation of required sequencing adapters. These potential issues must be considered when preparing isolated plasmid DNA for library preparation and sequencing.

There is a gap in the literature for detailed methods describing how to extract and prepare pDNA for library preparation and sequencing on 2nd and 3rd generation DNA

sequencers. Some recent publications describe the use of the 3rd generation sequencer, the Oxford Nanopore MinION, for obtaining long sequence reads for isolated, large plasmids (Szabó *et al.*, 2016; Li *et al.*, 2018); however the methods described are very brief and do not take into account the unique characteristics of plasmids that could potentially prevent them from being sequenced properly. Much of the literature on preparing isolated pDNA for sequencing is geared towards preparing plasmids, cosmids, and bacterial artificial chromosomes (BACs) for 1st generation Sanger sequencing, and oftentimes these methods rely upon the use of a commercial extraction kit. Williams *et al.* describe their method for "facile recovery of individual high molecular weight, low-copy plasmids for sequencing," however the protocol can take 2-3 days to perform and involves labor-intensive protocols such as extracting bands from agarose gels as a method to separate pDNA from chromosomal DNA (Williams *et al.*, 2006).

We present here a simple, rapid, and efficient method for extracting and preparing pure plasmid DNA in sufficient quantity and quality for sequencing library preparation on two sequencing platforms. Our large plasmid extraction technique is alkaline lysis based, does not require the use of a potentially expensive commercial extraction kit, and can be performed in approximately 2 hours using reagents and equipment that are common to molecular biology-equipped laboratories. We discuss the challenges faced when working with plasmid DNA, provide the specific techniques used to overcome these challenges and show the results of multiplexed Illumina MiniSeq and Oxford Nanopore MinION sequencing runs and the data produced as a result of our efforts. We hope that these techniques will prove useful to those interested in sequencing single, large plasmids (up to potentially 300kb or larger) on either Illumina or Oxford Nanopore

Technologies (ONT) sequencers for the purpose of plasmid genome assembly and characterization.

Methods

Plasmid extraction. Tetracycline resistant cells were cultured at 37 ºC overnight in 10 mL of TSB amended with tetracycline (12.5 µg/mL) (Figure 1). A 1.5 mL aliquot of turbid broth was centrifuged at 10,000 x g for 5 min and the supernatant was removed from the cell pellet by aspiration. The cell pellet was resuspended using 100 µL of an alkaline resuspension buffer (10 mM EDTA; 50 mM dextrose; 10 mM Tris-Cl, pH 8.0). One hundred microliters of freshly prepared 0.2 M NaOH/1% SDS were added to lyse the cells. Cells and SDS solution were mixed by inversion 5X and allowed to set at room temperature for 5 min. One hundred and fifty microliters of ammonium acetate (7.5 M) were added to reduce pH, immediately followed by the addition of 150 µL of chloroform, to denature proteins. The tube was gently inverted 5X before placing it on ice for 10 min to allow the plasmid DNA to dissolve. The mixture was centrifuged at 16,000 x g for 10 min and the supernatant was removed by aspiration and added to 200 µL PEG/NaCl solution (30% polyethylene glycol 8000, 1.5 M NaCl) to aid in the separation of plasmid from chromosomal DNA. This was chilled on ice for 15 min and centrifuged at 16,000 x g for 10 min. The supernatant was removed by aspiration. One milliliter of freshly prepared, chilled 70% ethanol was added to the tube and centrifuged at 10,000 x g for 5 min. The supernatant was removed and the tubes were allowed to air dry for no longer than 10 min. After drying, the remaining plasmid DNA pellet was reconstituted in 50-100 µL of MilliQ ddH_2O (MQ water). Plasmid DNA was allowed to dissolve for >24 hrs at 4°C and then stored for long-term at -20°C. RNAse A (Amresco, Solon, OH) was added to a concentration of 10 µg/mL to degrade RNA.

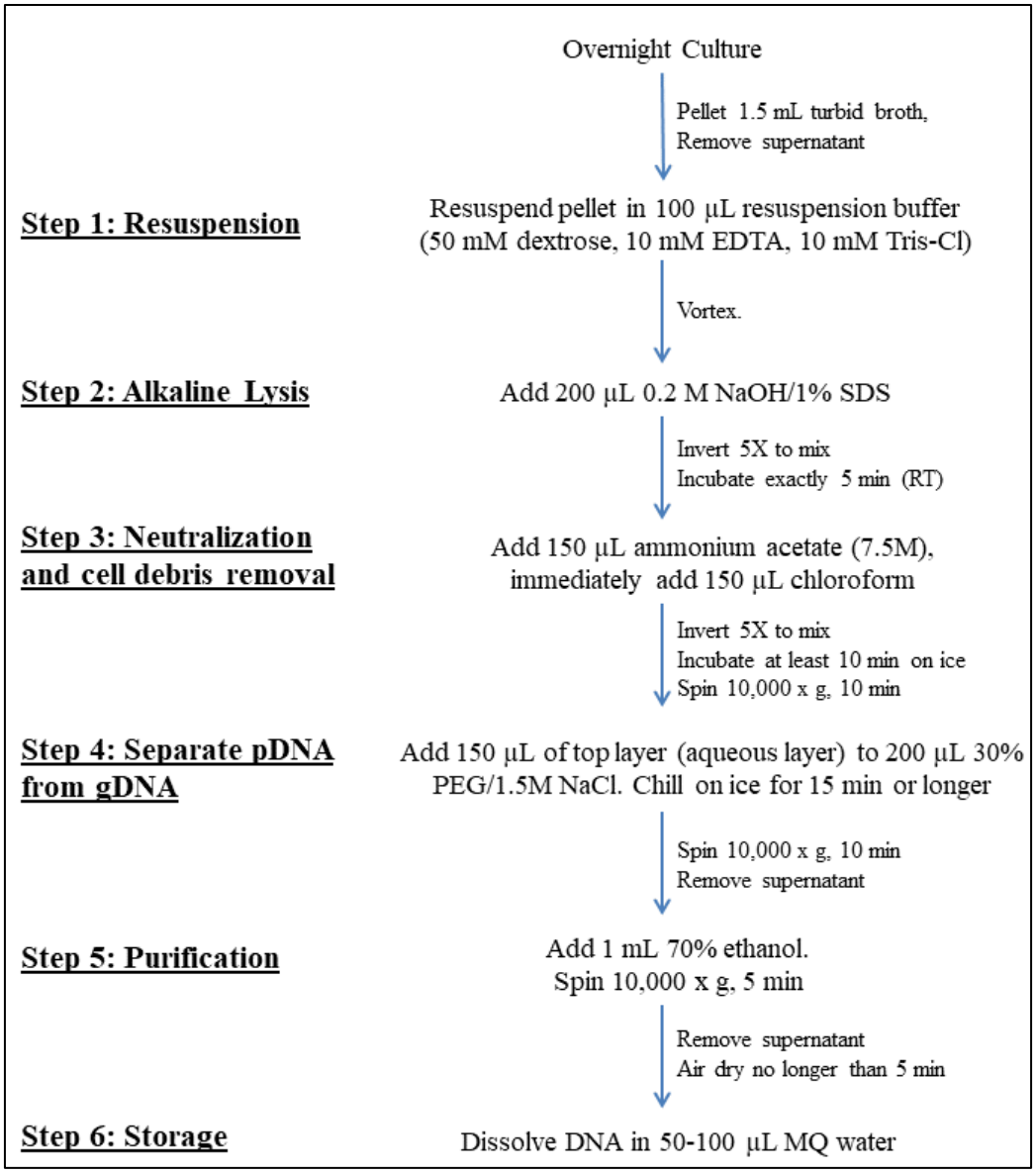

Figure 1. Plasmid extraction workflow.

Scaled-up plasmid extraction. To obtain larger quantities of DNA that are required for MinION sequencing (1-1.5 µg total DNA), the plasmid isolation protocol was scaled up using larger volumes of *E. coli* cultures (Libuit, 2016). Using a 50 ml Falcon® tube, EC100 cells harboring captured tet^R plasmids were grown in 30 mL of trypticase soy broth with shaking at 37º C overnight (180 RPM). Cells were harvested by centrifugation at 10,000 x g for 5 min and resuspended in 2 mL of resuspension buffer

(10 mM EDTA; 50 mM dextrose; 10 mM Tris-Cl, pH 8.0). Four milliliters of 0.2 M NaOH/1% SDS were added and the mixture kept at room temperature for 5 min. Three milliliters of ammonium acetate (7.5 M) and 3 mL of chloroform were added and the tube was gently swirled 5X before placing it on ice for 10 min. The tube was then centrifuged at 16,000 x g for 10 min. The supernatant containing plasmid DNA was added to 4 mL PEG/NaCl solution and chilled on ice for 15 min. Plasmid DNA was pelleted by centrifugation at 16,000 x g and resuspended in 1 mL of MQ water. DNA samples were kept at 4º C for >24 hrs to allow plasmid DNA to dissolve and then the samples were stored at -20°C.

The presence of plasmid DNA (pDNA) was confirmed by agarose gel electrophoresis. A 1% agarose gel was cast by boiling 0.6 g of agarose in 60 mL of 1X TAE buffer (40 mM Tris-acetate, 1 mM EDTA, pH 8.0). Eight microliters of plasmid DNA were mixed with 2 µL of 5X loading dye and the mixture run on the gel at 3.5V/cm for 90 min. In ladder lanes, 2 µL of 5X loading dye were mixed with 1 µL of lambda/HindIII digest and 7 µL of 1X TAE. Gels were stained in a 3X GelRed Nucleic Acid Gel Stain (Biotium, Hayward, CA) bath for at least 20 min and rinsed in DI water for 5 min. Gel images were taken with a Bio-Rad ChemiDoc XRS+ System and analyzed using Image Lab software (Bio-Rad, Hercules, CA).

Pre-library preparation plasmid DNA treatment. To remove contaminating chromosomal DNA, plasmid DNA samples were treated with Plasmid-Safe™ ATP-dependent DNase (Epicentre Technologies, Madison, WI). Forty-two microliters of plasmid DNA were mixed with 5 µL of Plasmid-Safe™ 10X Reaction buffer (330 mM Tris-acetate (pH 7.5), 660 mM potassium acetate, 100 mM magnesium acetate, and 5

mM DTT), 2 μL of 25 mM ATP, and 1 μL of Plasmid-Safe™ DNase. For treating volumes of plasmid DNA greater than 42 μL, the volumes of the components listed above were increased proportionately. The tubes were incubated at 37°C for 30 min and the enzyme was inactivated by incubation at 70°C for 30 min.

Restriction enzyme digestion of plasmid DNA for sequencing. To linearize the plasmid DNA samples for Illumina MiniSeq sequencing, the restriction enzyme EcoRI (New England BioLabs, Ipswich, MA [NEB]), was used. Sixteen and a half microliters of plasmid DNA were mixed with 2 μL of NEBuffer™ EcoRI (NEB) and 0.5 μL of EcoRI. The tubes were incubated at 37°C for 30 min and the enzyme was inactivated by incubation at 70°C for 30 min. For treating volumes of plasmid DNA greater than 16.5 μL, the volumes of the components listed above were increased, proportionately.

To linearize the plasmid DNA samples for Oxford Nanopore MinION sequencing, the restriction enzyme Sau3AI (NEB) was used to partially digest the DNA. One-hundred and seventy-nine microliters of plasmid DNA were mixed with 20 μL of NEBuffer 1.1 (NEB) and 0.2 μL of Sau3AI (5000 U/μL) (NEB). Immediately after adding the enzyme, the tubes were flicked to homogenize the solution and briefly centrifuged. Directly following centrifugation, the tubes were placed into a pre-warmed 70°C heating block for 30 min to inactivate the enzyme.

After Plasmid-Safe™ and restriction digestion (with either EcoRI or Sau3AI) were performed, samples were purified and concentrated using 1X volume of Agencourt AMPure XP Beads according to the manufacturer's instructions (Beckman Coulter, Brea, CA).

Library preparation. Plasmid DNA purity was verified using a Synergy H1 Multi-Mode Reader (BioTek Instruments, Winooski, VT). DNA was quantified using a Qubit™ 2.0 fluorometer (Invitrogen, Carlsbad, CA) using either the Qubit dsDNA BR kit (Broad Range; quantitation range = 2 – 1000 ng total DNA) or the Qubit dsDNA HS kit (High Sensitivity; quantitation range = 0.2 - 100 ng total DNA) according to the manufacturer's instructions.

Library preparation for Illumina MiniSeq sequencing. Sixteen MiniSeq sequencing libraries were prepared (only 4 of 16 were plasmid samples, other samples were for unrelated projects) using 1 ng of digested (RNAse A, PS, and EcoRI) plasmid DNA (5 µL @ 0.2 ng/µL) per sample. The Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA) and the Nextera XT Index Kit (Illumina) were used to prepare and dual-index each of the libraries, respectively.

Five microliters of each digested pDNA sample were added to 10 µL of Tagment DNA buffer (TD, Illumina) in a hard-shell skirted 96-well PCR plate (Bio-Rad, Hercules, CA). Five microliters of Amplicon Tagment Mix (ATM, Illumina) were added to each well and mixed by pipetting up and down briefly. The 96-well plate was centrifuged at 280 x g for 1 min at room temperature. The 96-well plate was placed in a thermocycler and incubated at 55°C for 5 min and cooled to 10°C. Immediately following, 5 µL of Neutralize Tagment buffer (NT, Illumina) were added to stop the transposome reaction. The 96-well plate was centrifuged at 280 x g for 1 min and incubated at room temperature for 5 min.

Tubes containing Nextera XT dual index primers were arranged in a TruSeq Index Plate Fixture (Illumina) to assist with adding unique combinations of indices to the

samples. Across the top row of the plate fixture, tubes of the following i7 indices were arranged from left to right: N701 (5'-TCGCCTTA-3'), N702 (5'-CTAGTACG-3'), N703 (5'-TTCTGCCT -3'), N704 (5'-GCTCAGGA -3'). In the first column of the plate fixture, tubes of the following i5 indices were arranged from top to bottom: N517 (5'-GCGTAAGA-3'), N502 (5'-CTCTCTAT -3'), N503 (5'-TATCCTCT-3'), N504 (5'-AGAGTAGA -3'). A multi-channel pipette was used to transfer 5 µL of the i7 indices to each of the rows of wells containing samples. Five microliters of each of the i5 indices were transferred to each of the columns of wells containing samples. Fifteen microliters of Nextera PCR Master Mix (NPM, Illumina) were added to each well containing a sample and mixed by pipetting. The 96-well plate was centrifuged at 280 x g for 1 min. The indices were ligated to the tagmented DNA fragments and amplified using the following limited-cycle PCR program: 72°C for 3 min, 95°C for 30 s, 12 cycles of [95°C for 10 s, 55°C for 30 s, 72°C for 30 s], 72°C for 5 min, and hold at 10°C.

Following amplification/ligation, the libraries were purified using Agencourt AMPure XP magnetic beads (Beckman-Coulter). The 96-well plate was centrifuged at 280 x g for 1 min. Fifty microliters from each sample were transferred to a 0.8 mL 96-well midi plate (Fisher Scientific, Waltham, MA). Thirty microliters of AMPure XP beads (0.6X volume) were added to each of the wells and the plate was placed in a plate shaker for 2 min at 1800 RPM. The 96-well plate was incubated at room temperature for 5 min and then placed on a magnetic stand for 2 min. Once the beads pelleted and the samples cleared in color, the supernatant was removed by aspiration and discarded. The pellets were washed twice by adding 200 µL of freshly prepared, chilled 80% ethanol, incubating the 96-well plate on the magnetic stand for 30 s, and removing the supernatant

from each well. Residual ethanol was removed by aspiration and the 96-well plate was allowed to air dry for no longer than 10 min. The 96-well plate was removed from the magnetic stand and 52.5 µL of Resuspension Buffer (RSB, Illumina) were added to each of the wells. The 96-well plate was then placed in the plate shaker for 2 min at 1800 RPM and incubated at room temperature for 2 min to elute the libraries off of the beads. The 96-well plate was placed on a magnetic stand for 2 min and 50 µL of each of the samples were transferred to a new 96-well PCR plate.

Twenty microliters of each sample were transferred to a new 96-well midi plate for bead-based normalization of the libraries. Eight-hundred and twenty-five microliters of Library Normalization Additives 1 (LNA1, Illumina) and 150 µL Library Normalization Beads 1 (LNB1, Illumina) were added to a 15 mL Falcon™ tube and inverted 3X to mix. Forty-five microliters of the LNA1/LNB1 mixture were added to each well in the 96-well midi plate that contained libraries. The 96-well midi plate was placed in a plate shaker for 30 min at 1800 RPM and then on a magnetic stand for 2 min and the supernatant was removed and discarded. The libraries were washed by repeating the following steps twice: 45 µL of Library Normalization Wash 1 (LNW1, Illumina) were added to each well, the plate was placed in a plate shaker for 5 min at 1800 RPM, the plate was placed on a magnetic stand for 2 min, and the supernatant was removed and discarded. Following the wash steps, 30 µL 0.1N NaOH were added to each well. The plate was shaken at 1800 RPM for 5 min.

A new 96-well hard-shell PCR plate was labeled SGP (Storage Plate) and 30 µL LNS1 (Library Normalization Storage buffer 1) were added to each well and the plate was set aside. After the 96-well midi plate finished shaking, samples were resuspended

by mixing with a pipette. The 96-well midi plate was shaken again at 1800 RPM for 5 min. The plate was incubated on a magnetic stand for 2 min and the supernatant was transferred to the SGP plate, which was subsequently centrifuged at 1000 x g for 1 min. Libraries were pooled by transferring 5 μL of each library to a 1.5 mL Lo-bind microcentrifuge tube (Eppendorf, Hamburg, Germany) labeled "PAL" (pooled amplicon libraries).

After libraries were pooled, they were diluted and denatured following the manufacturer's instructions (*Denature and Dilute Libraries Guide for MiniSeq System,* rev Jan. 2016). Three microliters from the PAL tube were transferred to a 1.5 mL tube containing 997 μL of prepared Hybridization Buffer (Illumina). The tube was vortexed and briefly centrifuged. Two-hundred and fifty microliters were transferred to a new 1.5 mL tube. Two-hundred and fifty microliters of Hybridization Buffer were added and the tube was vortexed and briefly centrifuged. The tube was then placed on a heating block for 2 min at 98°C and then immediately on ice for 5 min.

A PhiX control library was spiked into the pooled libraries, as recommended by Illumina, which can aid in assessing the quality of the sequencing run and troubleshoot potential issues (e.g. unsuccessful library prep). A 10 nM PhiX stock library (Illumina) was thawed and 10 μL were combined with 15 μL of RSB resulting in a 4 mM PhiX library. This tube was vortexed and centrifuged briefly. The library was denatured by combining 5 μL of the 4 mM PhiX library and 5 μL of 0.1N NaOH in a new tube, which was vortexed and centrifuged briefly. The tube was incubated at room temperature for 5 min and 5 μL 200 mM Tris-Cl (pH 7) were added to the tube, vortexed, and centrifuged briefly. Nine-hundred and eighty-five microliters of chilled hybridization buffer were

added to the tube of denatured PhiX library resulting in a 20 pM PhiX library. The library was diluted further by combining 45 µL of 20 pM PhiX library and 455 µL of hybridization buffer in a new tube, which was vortexed and centrifuged briefly, resulting in a 1.8 pM PhiX library. Twenty-five microliters of the 1.8 pM PhiX library were combined with 475 µL of the diluted, pooled libraries to aim for a 5% PhiX library spike-in rate.

Pooled libraries (including PhiX spike-in) were loaded into a MiniSeq High Output Reagent Cartridge (300-cycles) and sequenced on an Illumina MiniSeq for a total of 26 hrs. Following the sequencing run, raw image data files were translated into nucleotides (basecalled), de-multiplexed (i.e. separated according to barcodes), and adapter trimmed using the built-in Generate FASTQ analysis workflow to produce paired-end reads.

Library preparation for Oxford Nanopore MinION sequencing. Six MinION libraries were prepared using 1 to 1.5 µg of digested (RNAse A, PS, and partial Sau3AI digested) plasmid DNA (53.5 µL @ 18.7 to 28 ng/µL) per sample (with the exception of one sample being a genomic DNA sample for an unrelated project). The 1D Ligation Sequencing Kit (SQK-LSK108, ONT, Oxford, UK) and 1D Native Barcoding Kit (PCR-free) (EXP-NBD103, ONT) were used to prepare and barcode the libraries. Care was taken throughout the library prep to minimize the amount of pipetting done in order to limit the amount of shearing of the DNA.

Prior to MinION library preparation, the optional FFPE DNA repair kit (NEB) was used to repair nicks in the DNA molecules. Fifty-three and a half microliters of each sample (at 18.7 - 28 ng/µL) were mixed with 6.5 µL FFPE repair buffer and 2 µL of

FFPE repair enzyme mix. For plasmids pCCP1 and pCCP2, the volumes of the FFPE repair buffer and enzyme mix were increased proportionately to account for the larger volume of pDNA samples. Tubes were mixed by gentle flicking and centrifuged briefly, then incubated at room temperature for 15 min. Sixty-two microliters of Agencourt AMPure XP beads (1X volume) were added to each sample and incubated at room temp for a minimum of 5 min. Tubes were briefly centrifuged and placed on a magnetic rack to pellet the beads for at least 2 min. The supernatant was removed and discarded. Two hundred microliters of freshly prepared, chilled 70% ethanol were added to each tube, with care taken not to disturb the pellet. Ethanol was removed and the ethanol wash was repeated once more. After removing the ethanol, tubes were air dried for no longer than 5 min, removed from the magnetic rack, and 47 μL of MQ water were added to elute the DNA off the beads. Tubes were gently flicked to resuspend the beads and the tubes were incubated at room temp for at least 2 min and placed on the magnetic rack for 2 min until the tubes cleared in color. Forty-six microliters were removed from each tube, with care taken not to carry over beads, and transferred to new 1.5 mL Eppendorf Lo-bind tubes.

One microliter was removed from each sample and quantified on a Qubit 2.0 fluorimeter using the dsDNA BR assay kit according to the manufacturer's instructions (Invitrogen). The DNA was then subjected to an end-repair and dA-tailing reaction. Forty-five microliters of FFPE-repaired DNA samples were mixed with 7 μL of Ultra II end-prep reaction buffer (NEB), 3 μL of Ultra II End-prep enzyme mix (NEB), and 5 μL MQ water. Tubes were gently flicked to mix and centrifuged briefly. Tubes were incubated at room temperature for 5 min then at 65ºC for 5 min. Tubes were centrifuged briefly, 60 μL of AMPure XP beads were added to each tube, and mixed by gentle

flicking. Tubes were incubated at room temperature for at least 5 min to allow DNA to bind to the beads and centrifuged briefly. Tubes were then placed on magnetic rack for at least 2 min and the supernatant removed and discarded. Two-hundred microliters of freshly prepared, chilled 70% ethanol were added to each tube, with care taken not to disturb the pellet. Ethanol was removed, discarded, and the ethanol wash was repeated once more. After residual ethanol was removed and tubes air dried, the tubes were removed from the magnetic rack. Twenty-six microliters of MQ water were added to elute the DNA. Tubes were gently flicked to mix and incubated for 5 min at room temperature, then placed on the rack for 2 min, allowed to clear, and 26 µL of each sample were transferred to new 1.5 mL Lo-bind tubes. One microliter of each sample was quantified on the Qubit using the dsDNA BR assay kit according to the manufacturer's instructions (Invitrogen).

Barcodes were assigned to each sample as follows: BC01=pEG1-06, BC02 =pEG1-06 (duplicate of same sample), BC03=*Staphylococcus* gDNA (unrelated project), BC04=pCCP1, BC05=pCCP2, BC06=pCCRT11-6. Five-hundred nanograms of each sample were diluted with MQ water to a volume of 22.5 µL. Two and a half microliters of a native barcode, BC01-BC06, (ONT) were added to each tube, followed by 25 µL Blunt/TA Ligase Master Mix (NEB). Tubes were gently flicked to mix and centrifuged briefly. Tubes were incubated for 10 min at room temperature. Fifty microliters of AMPure XP beads (1X volume) were added to each tube, gently flicked to mix, and centrifuged briefly. Tubes were incubated at room temperature for at least 5 min and then placed on a magnetic rack for 2 min. Supernatants were removed and discarded and 200 µL of fresh, chilled 70% ethanol were added to each tube. Ethanol was removed and

another 200 µL of 70% ethanol were added and subsequently removed. Residual ethanol was removed and tubes air dried for no longer than 5 min. Tubes were removed from the magnetic rack and 27 µL of MQ water were added to the tubes to elute the DNA. Tubes were gently flicked to mix before placing them back on the magnetic rack for 2 min. Twenty-six microliters of each sample were transferred to new 1.5 mL Lo-bind tubes, with care taken not to carryover beads. One microliter was removed from each sample and quantified on the Qubit using the dsDNA BR assay kit according to the manufacturer's instructions.

Samples were pooled in equimolar amounts to a total of 700 ng and diluted with MQ water to a total volume of 50 µL. Twenty microliters of BAM (Barcode Adapter Mix, ONT) were added to the tube, followed by 20 µL Quick Ligation Reaction Buffer (5X) (NEB), and 10 µL of Quick T4 DNA Ligase (NEB). The tube was gently flicked to mix, centrifuged briefly, and then incubated for 10 min at room temperature before adding 40 µL of AMPure XP beads. The tube was gently flicked to mix and incubated at room temperature for 5 min, then placed on a magnetic rack for 2 min and the supernatant removed. One-hundred and forty microliters of ABB (Adapter Bead Bind buffer, ONT) were added to the tube and beads were resuspended by gently flicking the tube. The tube was placed back on the magnetic rack and incubated for at least 2 min for the beads to pellet. Supernatant was removed and another 140 µL of ABB were added. Beads were resuspended by removing the tube from the rack and gently flicking to mix. Then beads were pelleted on the magnetic rack, and the supernatant was removed. The tube was removed from the magnetic rack, 15 µL of elution buffer (EB, ONT) were added, gently flicked to mix, and the tube incubated for 10 min at room temperature. The

tube was placed on the magnetic rack and allowed to pellet for at least 2 min. Fifteen microliters of the pooled libraries were transferred to a new 1.5 mL Lo-bind tube. This pooled library was stored on ice until loading into the flowcell.

MinION flowcell priming and loading. Prior to loading the library, an R9.4 SpotON Flowcell (FLO-MIN106, ONT) was primed according to the manufacturer's instructions. The priming port cover was slid open and a P1000 pipette was used to draw back a small amount (~20-30 µL) of buffer, being careful not to introduce any air into the flowcell. Flowcell priming mix was prepared by mixing 524 µL RBF (Running Buffer with Fuel Mix, ONT) with 624 µL of MQ water. Eight hundred microliters of priming mix were slowly loaded into the flowcell via the priming port, with care being taken not to introduce air bubbles. After 5 min, the SpotON port was removed and 200 µL of priming mix were added via the priming port. In a new Lo-bind tube, 35 µL RBF were mixed with 2.5 µL MQ water, 25.5 µL LLB (Library Loading Beads, ONT), and 12 µL of the pooled library. Using a P200 pipette, the library was then gently mixed by pipetting and loaded to the SpotON port in a dropwise fashion, ensuring each drop flowed into the flowcell prior to the next drop falling.

A sequencing run was set up using MinKNOW software (v. 1.10.16) using the built-in python script "NC_48Hr_sequencing_FLO-MMIN106_SQK-LSK108_plus_basecaller.py" (protocols v. 1.10.11.1). After 8 hrs of sequencing, 18.75 µL of the prepared library (8.75 µL RBF + 0.6 µL MQ water + 6.4 µL LLB + 3 µL pooled library) were added to the flowcell using the technique described above. MinION sequencing is typically carried out for 48 hours to maximize the data yield from the

flowcell, however the run was stopped after approximately 16 hours because the nanopores on the flowcell had degraded and stopped producing data.

Results

Plasmid extraction. Plasmids were extracted using a procedure developed in our laboratory for the isolation of large, low-copy plasmids (Gehr, 2013). The protocol is alkaline lysis-based and was used for isolating plasmids ranging in size from approximately 58-121 kb. The extraction results in pure, RNA-free plasmid DNA (pDNA) as indicated by the OD260/280 and OD260/230 ratios measured using a Synergy H1 Multi-Mode Reader. All pDNA samples used for downstream applications had OD260/280 ratios between 1.8-2.0 and OD260/230 ratios of 2.0-2.2, which is generally accepted as pure for DNA samples (Sambrook, 2006).

Pre-library preparation plasmid DNA treatment. Prior to the sequencing of plasmid DNA samples on both the MiniSeq and MinION, steps were taken to remove RNA and chromosomal DNA as well as to linearize circular pDNA molecules. RNA was removed using RNase A and chromosomal DNA was removed using Plasmid-Safe DNase (PS), an exonuclease designed to digest any non-circular DNA, and was verified through gel electrophoresis (Figure 2). Lanes 2 and 3 were loaded with pEG1-06 DNA that was extracted from electroporated (EC100) cells using the plasmid extraction protocol and treated with RNAse A. RNA was not observed in either of the samples. The sample in lane 3 was additionally treated with PS and two bands were observed. The lower band at approximately 23.1 kb is the supercoiled form of the plasmid and the other band (>23.1 kb) is the relaxed circle form of the plasmid (Figure 2).

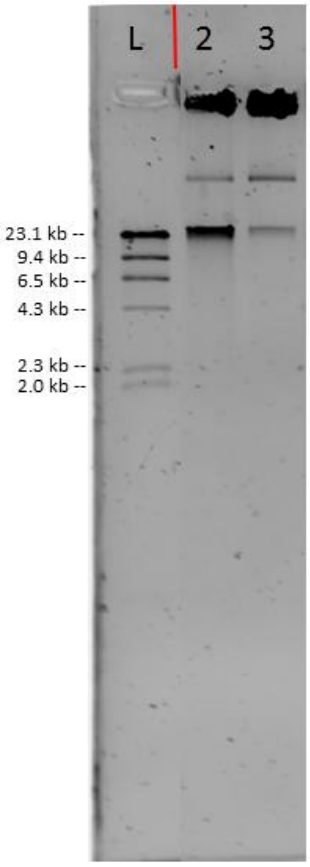

Figure 2. Plasmid DNA extracted from EC100 cells containing pEG1-06, visualized on a 1% agarose gel stained with GelRed. Lane "L" lane: lambda/HindIII digest. Lane 2: pDNA sample of pEG1-06, treated with RNAse. Lane 3: the same sample as lane 2, additionally treated with PS enzyme. Red line indicates where gel image was cut and edited to remove unnecessary gel lanes.

Linearization of the pDNA was necessary for two reasons. (1) If the pDNA molecules are in a supercoiled state, then the Qubit dsDNA-specific dye will not be able to efficiently bind to the pDNA, thus resulting in erroneous pDNA concentration readings (personal communication, ThermoFisher). (2) To successfully prepare pDNA for MinION sequencing, linear DNA molecules are required for the ligation of sequencing adapters on to the ends of the DNA fragments.

Restriction enzyme digestion of plasmid DNA for sequencing. Following plasmid extraction and treatment with RNAse A and PS, the plasmids were digested with

EcoRI and verified via agarose gel electrophoresis (Figure 3). Fragmentation was performed enzymatically using the restriction enzyme EcoRI prior to the Nextera XT DNA library preparation required for Illumina sequencing. The tagmentation reaction is the first step of the Nextera XT DNA library preparation which fragments DNA and ligates adapters to the fragments using a transposome. This reaction is very sensitive to DNA concentrations and can over-fragment or under-fragment if the initial quantity of DNA is higher or lower than the required 1 ng of DNA. Thus, to obtain accurate pDNA concentrations, we fragmented the pDNA prior to quantification with the Qubit™ fluorometer (Invitrogen, Thermo Fisher Scientific Inc).

Lane 2 contains pEG1-06 after treatment with PS enzyme and lane 3 contains the same sample as lane 2 but additionally digested by EcoRI (Figure 3). Lanes 4-5, 6-7, and 8-9 contain pCCP1, pCCP2, and pCCRT11-6 samples treated with PS and the same samples after EcoRI digestion, respectively. RNAse A was not added to the samples in lanes 2-5, and RNA can be seen in the bright patches below 564 bp. RNAse A was added to the samples in lane 6-9, and no RNA was observed. RNAse A was added to the samples in lane 3 and 5 prior to the MiniSeq library prep protocol.

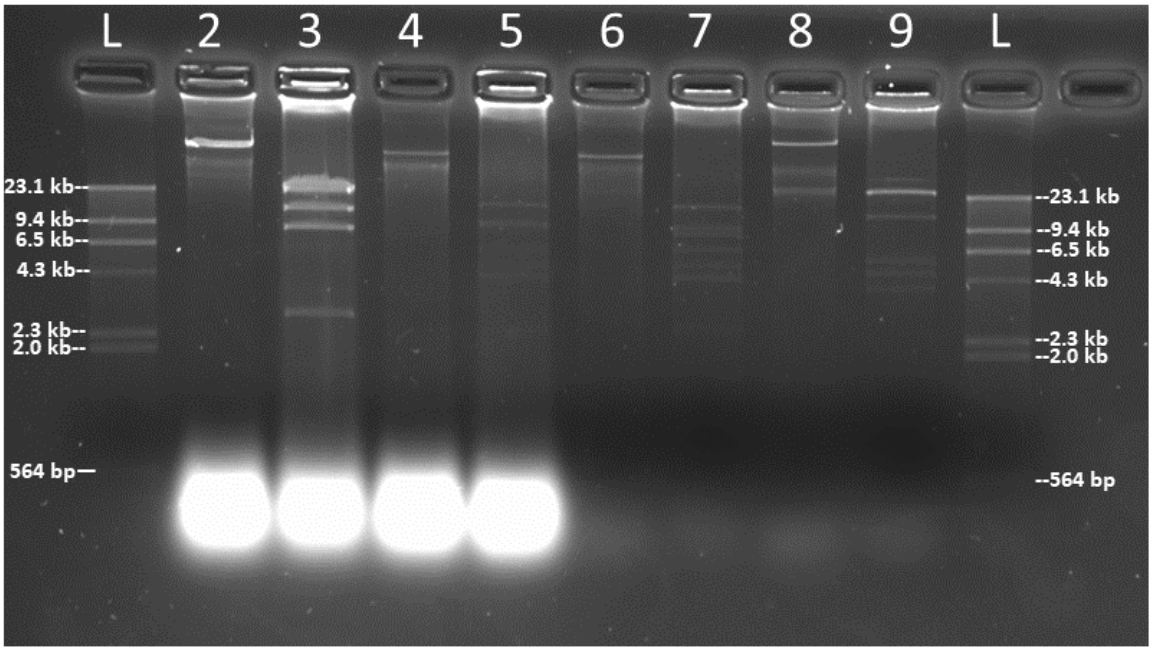

Figure 3. Plasmid DNA treated with Plasmid-Safe enzyme before (even lanes) and after (odd lanes) restriction digestion with EcoRI for Illumina MiniSeq library preparation. Samples treated with RNAse A in lanes 6-9. Lane L: lambda/HindIII digest marker, Lanes 2-3: pEG1-06, Lanes 4-5: pCCP1, Lanes 6-7: pCCP2, Lanes 8-9: pCCRT11-6.

Oxford Nanopore Technologies recommends that if one plans on performing the optional step of DNA fragmentation prior to the library preparation, that they use a G-tube™ (Covaris, Matthews, NC). The G-tube™ mechanically fragments the DNA by using centrifugal force to move the sample through a precisely manufactured orifice. Our previous attempts to fragment pDNA with a G-tube™, resulted in a MinION sequencing run that generated almost zero reads and it was hypothesized that the supercoiled state of pDNA did not allow successful fragmentation using the G-tube™, thus we opted to fragment the plasmids enzymatically prior to the library preparation (Libuit, 2016).

A “partial” digestion with Sau3AI, in this context, means that the step for incubating the tubes at 37°C was removed completely, to reduce the enzyme’s activity on the pDNA molecules. This was done in an attempt to fragment the pDNA as little as

possible, allowing for ligation of ONT sequencing adapters, but also to yield larger pDNA fragments than if a complete digestion was performed. Sau3AI was chosen because the enzyme recognizes a 4 bp restriction site (5'-GATC-3'), which will typically digest pDNA molecules more frequently than a 6-base cutter, so when the enzyme's activity is reduced, the partial digestion will yield a wide distribution of fragment sizes from 500bp up to the total size of the plasmid.

Following (scaled-up) plasmid extraction and treatment with PS and RNase A, the plasmids were partially digested with Sau3AI and visualized using agarose gel electrophoresis (Figure 4). Lane 2 contains pEG1-06 DNA treated with RNAse A and PS enzyme. Two bands can be seen, one at 23.1 kb that is the supercoiled form of the plasmid and the second, brighter band above 23.1 kb that is the relaxed circle form of the plasmid. Two bands can also be seen in lanes 4, 6, and 8 that are the supercoiled and relaxed circle forms of pCCP1, pCCP2, and pCCRT11-6, respectively. The relative sizes of the plasmids are reflected in lanes 2, 4, 6, and 8 as the relaxed circle bands for pCCP1 and pCCP2 (lanes 4 and 6, respectively) appear to migrate at the same (fastest) rate, while the relaxed circle band for pEG1-06 (lane 2) is higher up, and the relaxed circle band for pCCRT11-6 is highest.

Lane 3 contains the same pEG1-06 sample that is in lane 2, but partially digested with Sau3AI. A smear is observed from ca. 23 kb down (Figure 4). Lanes 4-5, 6-7, and 8-9 contain plasmids pCCP1, pCCP2, and pCCRT11-6 respectively, before and after partial digestion with Sau3AI (Figure 4). A smear is also observed in lanes 5, 7, and 9, however the smear is much lighter, indicating a lower quantity of pDNA present. They also appear

to have a lower average molecular weight, suggesting the plasmids were digested more frequently with Sau3AI.

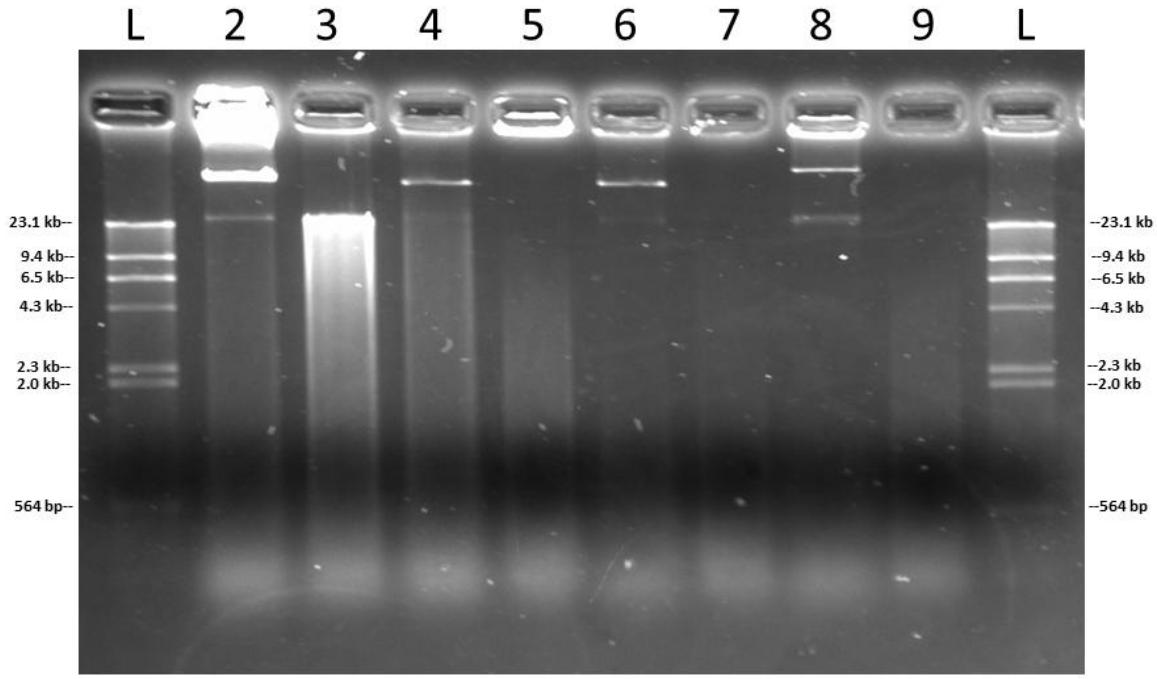

Figure 4. Plasmid DNA treated with Plasmid-Safe and RNAse A before (even lanes) and after (odd lanes) partial restriction digestion with Sau3AI for Oxford Nanopore library preparation. Lanes L: lambda/HindIII digest, lanes 2-3: pEG1-06, lanes 4-5: pCCP1, lanes 6-7: pCCP2, lanes 8-9: pCCRT11-6.

Library preparation. Once pDNA was extracted from EC100 cells and digested (RNAse A, PS, and either EcoRI or Sau3AI partial digestion), libraries were prepared for sequencing on an Illumina MiniSeq and an Oxford Nanopore MinION. Prior to library preparation, sample purity was verified to check for an OD 260/280 ratio between 1.8 and 2.0 and an OD 260/230 ratio between 2.0 and 2.2. For MiniSeq sequencing, digested plasmid DNA samples were diluted and quantified using the Qubit dsDNA HS kit (Table 1). Samples were then diluted further to 1 ng in 5 µL (0.2 ng/µL).

Table 1. pDNA concentrations prior to Illumina library preparation. Concentrations were measured using the Qubit fluorimeter.

Plasmid	DNA concentration (ng/µL)
pEG1-06	0.55
pCCP1	2.08
pCCP2	0.63
pCCRT11-6	1.31

For MinION sequencing, digested pDNA samples were concentrated using Agencourt AMPure XP beads and quantified using the Qubit dsDNA BR kit (Table 2). pEG1-06 replicate 2 was diluted to 1.5 µg in 53.5 µL (28 ng/µL). pEG1-06 replicate 1 and pCCRT11-6 samples were not diluted further. pCCP1 and pCCP2 samples had the minimum required quantity of DNA (1000 ng) but the DNA concentrations were too low. To mitigate this during the ONT library preparation, the volumes of reagents during first reaction (FFPE DNA repair) were increased proportionately to compensate for the higher volume of the pDNA samples.

Table 2. pDNA concentrations prior to Oxford Nanopore Library preparation. Concentrations were measured using the Qubit fluorimeter.

Plasmid	DNA concentration (ng/µL)
pEG1-06 replicate 1	23.4
pEG1-06 replicate 2	29.5
pCCP1	16.5
pCCP2	19.4
pCCRT11-6	27.0

MiniSeq sequencing run. Libraries were prepared using digested plasmid DNA and were sequenced on an Illumina MiniSeq sequencer. During the first few cycles of the sequencing run, single stranded DNA libraries hybridize with the flowcell and are clonally amplified while immobilized on the flowcell to produce “clusters” of clonally

amplified libraries. Clusters are then detected, mapped, and are subjected to a quality filter to remove the least reliable clusters from analysis (Illumina, 2016). Reads that are generated from the clusters that passed the quality filter are called "reads passed filter (PF)." Cluster density is a measure of signal intensity (emitted light) in a given area on the MiniSeq flowcell (resulting from clusters in close proximity) and the metric can influence run quality, reads passing filter, read quality scores, and total data output. For the MiniSeq the optimal range of cluster density set by Illumina is 170-220 K clusters/mm^2 and the cluster density from this sequencing run was 110 ± 2 K clusters/mm^2 indicating underclustering of the flowcell (Table 3).

Table 3. MiniSeq sequencing run statistics.

Total bases sequenced	5.51 Gb			
Total # of reads	18,544,771			
# of reads Passed Filter (PF)	17,546,094			
	pEG1-06	*pCCP1*	*pCCP2*	*pCCRT11-6*
	964,110	506,701	653,046	753,409
Percent of reads identified/sample (PF)	5.4947 %	2.8878 %	3.7218 %	4.2938 %
# of bases ≥ Q30 (PF) (%)	5370.1 Mb (96.36 %)			
Cluster density	110 ± 2 K clusters/mm^2			
% PhiX	11.98 %			

Another important metric for assessing the quality of a MiniSeq sequencing run is the percentage of reads that uniquely aligned to the PhiX reference genome. PhiX is often spiked into the pooled libraries as a control library to troubleshoot potential issues such as failed library preparations but also can be useful when there is low nucleotide "diversity" among samples (e.g. sequencing amplicons with high % GC content). The percentage of reads aligned to PhiX was 11.98%, which is higher than the expected 5%, but nonetheless acceptable (Illumina, 2016) (Table 3).

Phred quality score (or simply Q score) is the metric used for sequence read accuracy and has an inverse, logarithmic relation to the probability of error of a base call (Ewing and Green, 1998). For example, Q10 indicates the probability that 1 of every 10 base calls will be an error (10% error rate); Q20 indicates the probability that 1 in 100 base calls will be an error (1% error rate); Q30 indicates the probability that 1 in 1,000 base calls will be an error (0.1% error rate).

The MiniSeq sequencing run generated a total of 5.51 Gigabases (Gb) of sequence data in 18,544,771 paired-end (PE) reads with an average read length of 150 bp. 17,546,094 of these reads passed the filter and 96.36% of all bases sequenced had a Q score ≥ 30 (Table 3, Figure 5). Average Q scores across all nucleotide positions in all of the reads ranged from 32.85 - 36.62 (Figure 6).

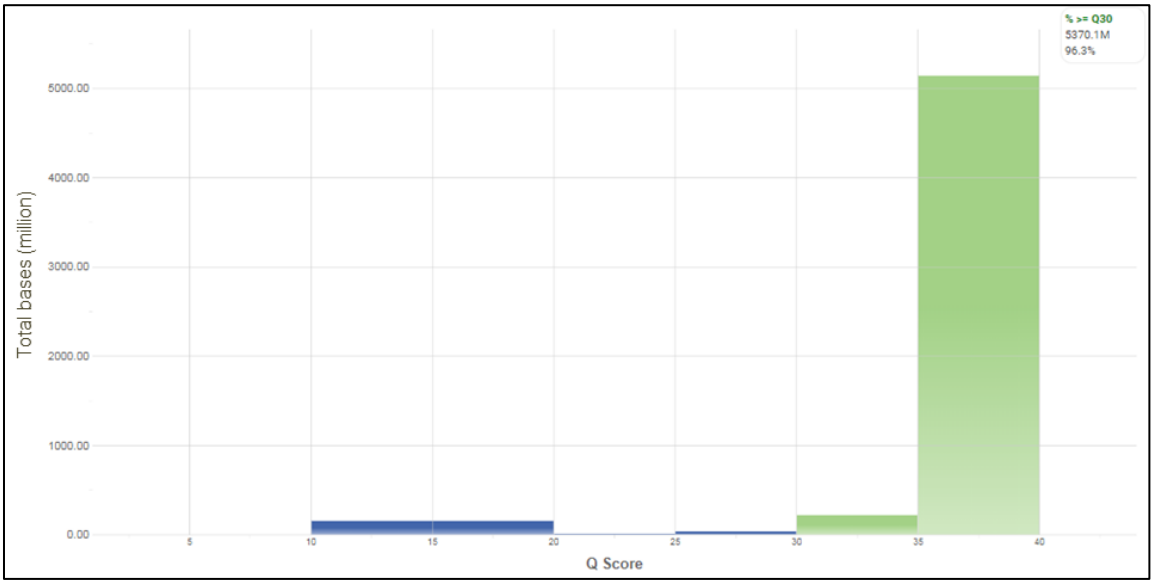

Figure 5. Distribution of Q scores for all bases sequenced during the MiniSeq run. Blue bars indicate bases with a Q score < 30 and green bars indicate bases with a Q score ≥ 30.

Reads were de-multiplexed and adapters and barcodes were removed using the Generate FASTQ analysis workflow. The pEG1-06 sample had the highest number of

reads (among the 4 plasmid samples, not including the 12 other samples) with 964,110 PE reads (5.4847% reads PF) (Table 3, Figure 7). The pCCRT11-6 sample had the second highest amount of reads with 753,409 PE reads (4.2938% reads PF). The pCCP2 and pCCP1 samples had the 3rd and 4th highest amount of reads with 653,046 (3.7218% reads PF) and 506,701 PE reads (2.8878% reads PF), respectively (Table 3, Figure 7).

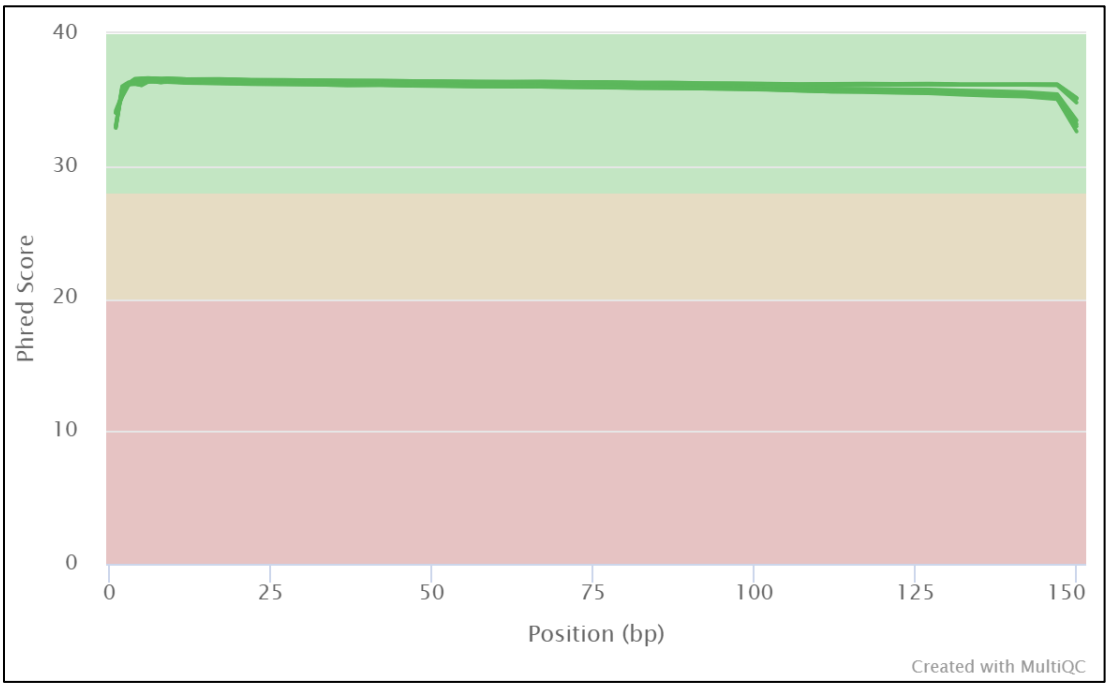

Figure 6. Average Q scores for each nucleotide position in the Illumina reads. Eight lines are plotted and each line represents either the forward or reverse reads for each of the 4 samples. Plot was generated using MultiQC (Ewels *et al.*, 2016).

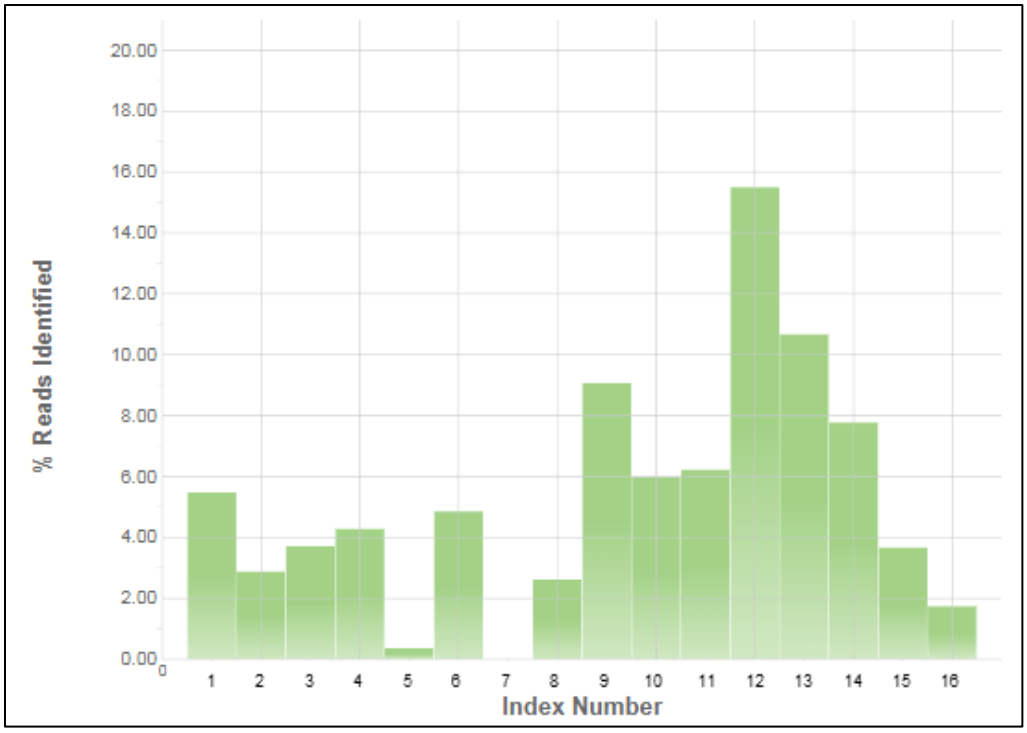

Figure 7. Percentage of reads identified for each index (barcode). Index assignments are as follows: index 1: pEG1-06, index 2: pCCP1, index 3: pCCP2, index 4: pCCRT11-6, indices 5-16: unrelated samples.

MinION sequencing run. Libraries were prepared using digested pDNA and were sequenced on a MinION sequencer. Nanopore sequencers measure picoampere (pA) level changes in ionic current as single stranded DNA (ssDNA) molecules pass through a nanopore, and these measurements are considered "raw signal trace" data (file type: fast5). When a molecule of ssDNA passes through a pore, the raw signal data is recorded and considered as a single 1D read. The 16 hr sequencing run generated 220,775 reads containing approximately 459 Mb of sequence data (Figure 8a; Table 4). After 16 hrs of sequencing, the nanopores present in the flowcell had degraded and the device stopped producing data (Figure 8a). The sequencing run was stopped and the raw signal data was basecalled into nucleotides using Albacore (ONT official basecalling software, https://community.nanoporetech.com/downloads) resulting in fastq files.

Table 4. MinION sequencing run general statistics. Statistics were calculated using Porechop and NanoPlot.

Total # of bases sequenced	459,280,985 bases				
# of reads	220,775				
	pEG1-06 replicate 1	*pEG1-06 replicate 2*	*pCCP1*	*pCCP2*	*pCCRT11-6*
	13,799	39,727	29,717	18,931	62,671
# of bases/sample	39,653,462	138,396,462	30,941,099	19,766,071	81,471,888
Mean read length (bases)	2,080				
Mean read length/sample (bases)	*pEG1-06 replicate 1*	*pEG1-06 replicate 2*	*pCCP1*	*pCCP2*	*pCCRT11-6*
	2,958.7	3,572.7	1,137	1,130.3	1,399.1
Median read length (bases)	1,093 bases				
Median read length/sample (bases)	*pEG1-06 replicate 1*	*pEG1-06 replicate 2*	*pCCP1*	*pCCP2*	*pCCRT11-6*
	1,550	1,972	914	909	998
Mean read Q score	9.1				
Median read Q score	9.5				
# reads >Q7 (%), Mbases	191,729 (86.8%), 410.2 Mb				
# reads >Q10 (%), Mbases	74,744 (33.9%), 175.8 Mb				
Longest read	127,065 bases				
Read N50	3,636 bases				
# of active channels	437 out of 512 channels				
Elapsed time of seq. run	16 hrs				

The mean read length for all the reads was 2,080 bases, the median read length was 1,093 bases, and the read N50 was 3,636 bases (Figure 8b, Table 4). Read N50 is

defined as the read length such that the reads of this length or greater sum to at least half of the total bases in all reads (Jain *et al.*, 2018).

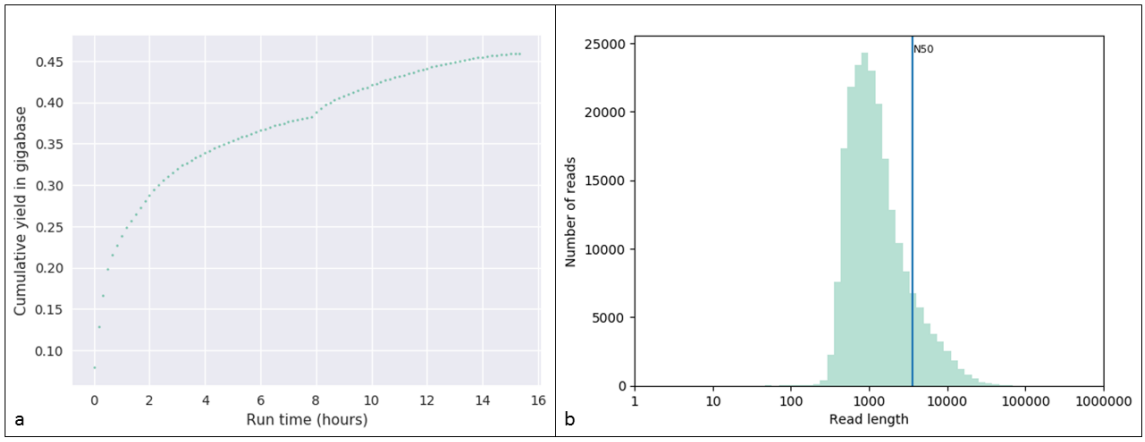

Figure 8. (a) Cumulative yield of sequence data over the length of sequencing run. The increase in slope at 8 hrs is due to the loading of additional library to the flowcell after 8 hours of sequencing. (b) Read length distribution with read N50 line at 3,636 bases. Both plots were generated using NanoPlot (De Coster, 2017).

Reads were de-multiplexed using both Albacore and Porechop (https://github.com/rrwick/porechop) and barcode sequences were removed from the reads using Porechop. The first pEG1-06 replicate had 13,799 reads containing a total of 39,653,462 bases with a mean read length of 2,958.7 bases and a median read length of 1,550 bases (Table 4, Figure S1). The second pEG1-06 replicate had 39,727 reads containing a total of 138,396,462 bases with a mean read length of 3,572.7 bases and a median read length of 1972 bases (Table 4, Figure S2). pCCP1 had 29,717 reads containing a total of 30,641,099 bases with a mean read length of 1,137 bases and a median read length of 914 bases (Table 4, Figure S3). pCCP2 had 18,931 reads containing a total of 19,766,071 bases with a mean read length of 1,130.3 bases and a median read length of 909 bases (Table 4, Figure S4). pCCRT11-6 had 62,671 reads

containing a total of 81,471,888 bases with a mean read length of 1,399.1 bases and a median read length of 998 bases (Table 4, Figure S5).

The mean read Q score for all ONT reads was 9.1 and the median read Q score was 9.5 (Figure 9, Table 4). 191,729 (86.8%) of the reads (which contain 410.2 Mb) had a Q score greater than 7 (Table 4). Q7 represents approximately a 19.9% error rate. 74,744 (33.9%) of the reads (which contain 175.8 Mb) had a Q score greater than 10. pEG1-06 replicate 1 reads had a mean Q score of 9.6 and a median Q score of 9.8 (Figure S6). pEG1-06 replicate 2 reads had a mean Q score of 9.5 and a median Q score of 9.7 (Figure S7). pCCP1 reads had a mean Q score of 9.6 and a median Q score of 9.8 (Figure S8). pCCP2 reads had a mean Q score of 9.6 and a median Q score of 9.7 (Figure S9). pCCRT11-6 reads had a mean Q score of 9.6 and a median Q score of 9.8 (Figure S10).

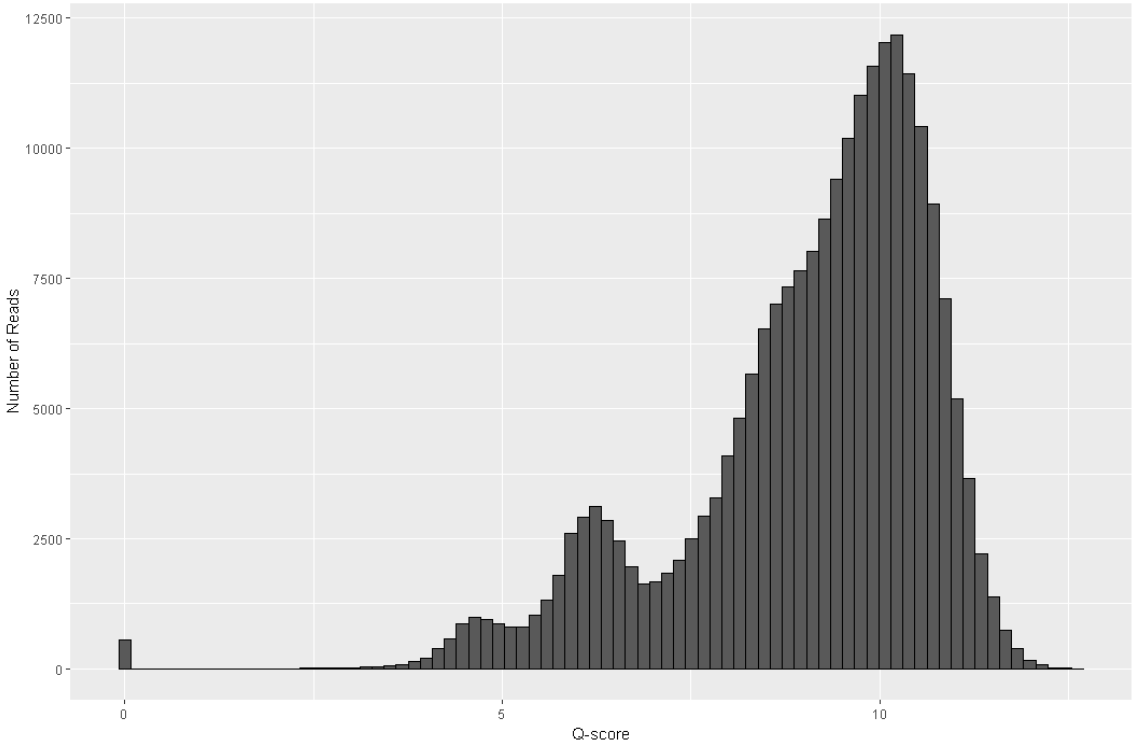

Figure 9. Mean read quality score distribution for all reads from the MinION sequencing run. The highest average quality score from a read was 12.62. Median quality score was 9.5, mean quality score of all reads was 9.1.

Discussion

Plasmid extraction and pre-sequencing treatment. In this study, we present a method for quickly and efficiently isolating plasmid DNA (pDNA) without the use of a commercial plasmid extraction kit, and methods for the treatment of pDNA samples prior to sequence library preparation for Illumina and Oxford Nanopore DNA sequencing. These methods include the isolation of pDNA by alkaline lysis, removal of chromosomal DNA by Plasmid-Safe digestion, enzymatic fragmentation of pDNA, accurate quantification of pDNA using fluorometric methods, and library preparation using the Nextera XT DNA library prep kit for Illumina sequencing and the PCR-free 1D Ligation library prep kit for ONT sequencing. We demonstrate the efficacy of this method by preparing and sequencing four large, tetracycline-resistance plasmids on both sequencing platforms.

Traditional plasmid sequencing strategies involve plasmid DNA purification and shotgun sequencing; however, purification of large plasmids (> 50 kb) is typically difficult, time-consuming, and can be expensive if commercial kits are used to extract the plasmids (Smalla *et al.*, 2015). Our method of plasmid extraction by alkaline lysis allows one to quickly obtain pure pDNA from plasmids up to ~122 kb in size (up to 300 kb, unpublished) from a broth culture in approximately 1-1.5 hrs. This technique yields pure pDNA samples that are free of RNA, protein, and salts that may inhibit downstream applications that are sensitive to contaminants (e.g. library preparation and DNA sequencing). This extraction technique has primarily been used in our lab to extract from *E. coli* and *Salmonella enterica* (Gram-negative bacteria) cells however this can also be

applied to extracting pDNA from Gram-positive genera such as *Staphylococcus*, *Enterococcus*, and *Bacillus*.

When the plasmid extraction technique was scaled up to extracting from larger volume cultures (e.g. 30 mL broth cultures), we were able to extract sufficient quantities of DNA for ONT sequencing, which require 400-1000 ng of total DNA (depending upon the type of ONT library prep kit used). It is often difficult to obtain sufficient quantities of plasmid DNA, especially if the plasmid is of low-copy number. One additional method to aid in increased pDNA yields (from plasmids expressing antibiotic resistance) is to amend the broth cultures with an antibiotic to ensure that all cells contain the plasmid of interest.

Researchers that aim to characterize the total plasmid content from a given environment the will often sequence the "meta-plasmidome" where expected pDNA yields are low after extraction, and because of the low yields, amplification strategies are typically employed to increase pDNA quantities prior to sequencing. Phi29 polymerase rolling circle amplification has been used to amplify circular DNA molecules (which inherently excludes chromosomal DNA) (Kav *et al.*, 2013). The disadvantage to this approach is that there are known biases with rolling circle amplification, as the amplification will increase the abundance of smaller plasmids over that of large plasmids (Norman *et al.*, 2014). This bias can be reduced, but at the cost of an extra step to separate plasmids based on their size prior to amplification.

Removal of chromosomal DNA. Our alkaline lysis-based plasmid extraction technique typically yields sufficient quantities of pDNA for many downstream applications; however there are still small amounts of chromosomal DNA which carry

over into the final sample. To remove the chromosomal DNA, we elected to use an exonuclease, Plasmid-Safe ATP-dependent DNase, to digest linear chromosomal DNA. This enzyme works well for digesting linear fragments of chromosomal DNA and this removal can then be visualized on a standard agarose gel (Figure 2). There are some disadvantages to this approach. Any linear plasmid forms will be digested and lost. Plasmid typically exist in three conformations: (1) supercoiled conformation, where the DNA is tightly coiled and occupies the smallest volume of space within cells, (2) relaxed circle conformation where one of the two strands of pDNA has been nicked, allowing for the uncoiling of the pDNA molecules, and (3) linear, where both pDNA strands have been broken and the plasmid is completely uncoiled. Also, Plasmid-Safe digestion does require some form of sample purification prior to sequencing, although if the DNA is being used downstream for sequencing, the DNA purification steps typically necessary for library preparation will be sufficient.

In this study, we followed the manufacturer's instructions of incubating the plasmid-safe reaction at 37ºC for 30 minutes, with the goal of removing all chromosomal DNA. After sequencing, it was clear that some *E. coli* chromosomal material did get sequenced along with our plasmids of interest due to the presence of contiguous sequences that aligned with near 100% identity to an *E. coli* reference genome. This may have confounded the plasmid genome assemblies and allowed for assembly errors to occur, but this will be addressed in the Chapter 2 Results and Discussion section. We therefore recommend extending the 37ºC incubation to 1.5 hrs or longer to ensure the complete digestion of linear chromosomal DNA. In addition to visual confirmation of chromosomal DNA removal on an agarose gel, 16S rRNA PCR can be performed on the

samples, post-Plasmid-Safe treatment, to see if any chromosomal fragments remain undigested (Kav *et al.*, 2013).

Enzymatic fragmentation of pDNA. Accurate quantification of DNA samples is critical for sequencing, especially when using the Nextera XT library preparation kit where successful preparation of sequencing libraries is highly dependent on the input quantity of DNA (Illumina, 2016). The tagmentation reaction is the first step of the Nextera library prep which fragments DNA and ligates adapters to the fragments using a transposome. This reaction is sensitive to DNA concentrations and can over-fragment or under-fragment if the initial quantity of DNA is higher or lower than the required 1 ng of DNA, leading to low quality sequence runs.

We opted to enzymatically fragment the plasmids prior to sequencing, allowing for accurate quantification using the Qubit fluorimeter (personal communication, ThermoFisher) and to allow for the ligation of ONT sequencing adapters during the 1D Ligation library preparation protocol. We used enzymatic fragmentation rather than the ONT recommended mechanical fragmentation with a Covaris G-tube, due to a failed plasmid sequencing run that was likely caused by unsuccessful shearing of pDNA molecules prior to the library preparation protocol. (Libuit, 2016). It was hypothesized that the supercoiling state of the plasmid, caused the pDNA to pass through the G-tubes' orifice without fragmenting the pDNA molecules. Without free ends of pDNA fragments, ONT sequencing adapters could not be ligated, thus the pDNA is not sequenceable. Enzymatic fragmentation of the four plasmids, pEG1-06, pCCRT11-6, pCCP1, and pCCP2 was performed using the restriction enzyme Sau3AI prior to ONT sequencing. The read lengths produced by ONT sequencers are highly dependent upon the size of the

input fragment lengths, allowing for long read lengths (in our case, up to the total size of the plasmids) (Jain *et al.*, 2018). The partial digestion of plasmids dictated the resulting read lengths and they varied depending on the plasmid. For pEG1-06 digested by Sau3AI and visualized on a gel, we observed a smear ranging from ca. 23 kb down (Figure 4, lane 3) and a wide distribution of read lengths were reflected in the ONT data with reads up to 61 kb and a mean read length of 3.2 kb (Supplementary Figure S2). Plasmid pCCP2 was digested more frequently than pEG1-06, and this was likely due to either a lower quantity of pDNA prior to the digest than the pEG1-06 sample or the presence of more Sau3AI restriction sites being present in the pCCP2 genome (Figure 4, lane 7). This plasmid resulted in the shortest ONT mean read lengths of 1.1 kb; the longest read was 37.1 kb (Supplementary Figure S4).

For plasmid sequencing experiments using a ONT sequencer, we recommend experimenting with various restriction enzymes, with the goal of fragmenting the plasmids as little as possible in order to take advantage of the long read lengths of the MinION. Despite fragmenting the plasmids to sizes smaller than we had hoped, the reads were long enough to allow for the complete assembly of the genomes (Chapter 2 Results and Discussion: Figure 15, p. 74). Mechanical fragmentation methods may be more consistent in achieving longer read lengths, though these protocols would have to be optimized for plasmids due to the tendency of plasmids to form supercoils, which can be resistant to mechanical shearing methods. Regardless of plasmid fragmentation technique, we recommend assessing fragment length distributions by visualization on an agarose gel or using a Bioanalyzer prior to ONT library preparation.

MiniSeq sequencing. The data produced by the MiniSeq sequencing run was of very high quality (96.3% of bases sequenced ≥ Q30) however the sequencing run yielded approximately 5.5 Gb of sequence data, which was lower than the theoretical maximum of that the MiniSeq is capable of producing - 7.5 Gb. This was likely the result of the underclustering of the flowcell that was observed for the sequencing run (Table 3). Underclustering of the flowcell was an indication that we underestimated the quantity of the pooled sequencing libraries that were loaded into the sequencer, therefore the total amount of sequence data produced was lower than the theoretical maximum, but this allowed for the data to be of very high quality (Phred Quality scores). Often observed with longer Illumina MiSeq reads (300 bp in length), there is a significant drop off in quality scores when the last ~50 bases are sequenced, however the data produced throughout the entire length of the sequencing run were ≥ Q32 (Figure 6) (Shendure and Ji, 2008). Despite the attempt to normalize the concentrations of libraries prior to pooling and loading of the flowcell, we observed an uneven amount of data produced by each unique libarary, ranging from near 0% of read identified for a barcode, up to ~15% of read identified for a single barcode (Figure 7). We believe this was due to an uneven quantity of DNA going into the Nextera XT library prep, or the different libraries hybridizing to the flowcell in different proportions, thus producing a varying amount of data for each sample.

Nanopore sequencing. The quality of the data produced by the MinION sequencing run was of typical for that of ONT data which is estimated to be 92% accuracy for 1D reads (Jain *et al.*, 2016). The reads produced by the MinION run had mean Q score 9.1 (~87.7% accuracy), median Q score 9.5 (~88.8% accuracy) (Table 4).

This MinION run produced approximately 0.459 Gb of sequence data, which is low compared to the typical output of 3-5 Gb of sequence data typically produced by MinION sequencing runs (Leggett and Clark, 2017) however this was not an issue due to the relatively small size of the plasmid genomes and the high average depth of coverage achieved for the assemblies. After assembling the plasmid genomes and aligning the ONT reads back to the assemblies, the lowest average depth of coverage among the four plasmids was high - at 105.8X for pCCP1 (Chapter 2: Results and Discussion, Table 7).

Chapter 2

Exogenous capture, sequencing, characterization, and genomic analysis of four antibiotic resistance plasmids captured from the stream sediments of agriculturally-impacted streams in the Shenandoah Valley

Introduction

Large, self-transmissible plasmids are key vectors in the transfer of resistance, catabolic, and other genes among bacteria native to environments such as streams and wetlands (Herrick *et al.*, 2014; Botts *et al,* 2017). The evolution of antibiotic resistance in particular is known to be powerfully affected by conjugative plasmid transfer due to the ability of plasmids to exchange mobile genetic elements such as transposons and insertion sequences that often contain antibiotic resistance genes (Smillie *et al.*, 2010). This is further driven by the ease in which some plasmids can be horizontally transferred into a broad host range of bacteria. The dissemination and evolution of antibiotic resistance is largely driven by conjugative plasmids, so it is necessary to identify and understand the characteristics and behavior that contribute to the spread of antibiotic resistance among bacteria.

Traditionally, plasmids have been classified using conventional PCR-based methods for typing based on backbone genes such as plasmid replication genes (replicon typing) and mobility typing (MOB). However, the use of whole genome sequencing to study plasmids *in silico* has prompted a shift in the field of plasmid biology towards classifying plasmids based on their entire genomes instead of a single or a few loci (Orlek *et al.*, 2017). While replicon typing is useful for identifying incompatibility groups of plasmids, whole (plasmid) genome sequencing offers a finer level of resolution and allows for the differentiation of plasmids based on slight differences in nucleotide or

amino acid sequence. For example, the differentiation of IncA/C_2 type one and type two variant group plasmids is based on single coding sequence (CDS) which differs only slightly in length due to insertions and deletions. This CDS is located between the *traA* gene and *dsbC* gene located within the hallmark Tra suite of genes involved in conjugal transfer (Harmer and Hall, 2015). IncA/C_2 type one variants contain the CDS that is 1,832 amino acids (AA) in length and IncA/C_2 type two variants contain the CDS that is 1,847 AA in length.

The advent of whole genome sequencing has increased our understanding of the extrachromosomal elements and they have been increasingly studied with the advent of 3rd generation long read sequencing technologies like Pacific Biosciences and Oxford Nanopore Technologies sequencers (Li *et al.*, 2018; Botts *et al.*, 2017; George *et al.*, 2017). Long reads have enabled the complete assembly of many plasmids that are difficult to assemble with short reads alone. During this study in which the sequencing, genome assembly, and analysis of four plasmids were performed, we found assemblies using short read sequence data alone was not enough to resolve the plasmid genomes completely, and required the addition of long reads to obtain complete, circular assemblies.

Here we present a study on the exogenous capture and characterization of four transmissible, tetracycline resistance plasmids, one 71 kb IncP-1β multi-drug resistance plasmid previously captured in 2006 (Gehr, 2013), two 59 kb IncP-9 tetracycline resistance plasmids captured from Cooks Creek Park, and one 121 kb IncA/C_2 tetracycline resistance plasmid captured from Cooks Creek Rt. 11. We describe the complete nucleotide sequences of each of the four novel plasmids and compare their

genomes to the closest related plasmids to gain insight into their genes, predicted functions, and how they compare to plasmids of the same incompatibility groups.

Methods

Sampling sites. Stream sediment samples were collected from four streams in the Shenandoah Valley of Virginia: Muddy Creek, Cooks Creek, Pleasant Run and War Branch (Figure 10). Muddy Creek runs through Hinton, Virginia originating in the Allegheny Mountains and the sampling location is directly adjacent to a poultry processing plant. Muddy Creek, Cooks Creek, and Pleasant Run were chosen due to their historically high *E. coli* and coliform counts and the fact that they drain numerous cattle pastures and poultry farms. War Branch is a tributary of Muddy Creek and is higher in elevation, with no obvious agricultural impact.

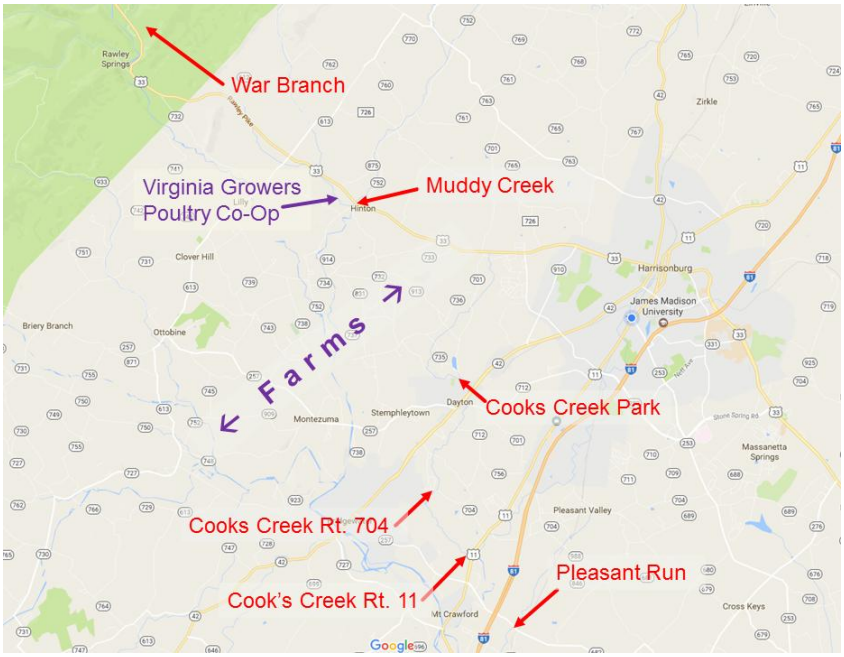

Figure 10. Locations of the stream sediment sampling sites.

Sample collection. Sediment samples were collected in sterile 50 mL Falcon tubes. The top layer of sediment was brushed away using a sterile spatula and the Falcon® tube was used to scoop sediment into the tube. Samples were transported to the lab on ice, stored at 4°C and processed within 24 hours.

Bacterial strains. *Escherichia coli* strain LA61 was used as the recipient in exogenous plasmid captures (Gehr, 2013). LA61 is a plasmid-free, tetracycline-sensitive (tet^S), rifampin-resistant (rif^R), strain previously isolated in our laboratory from Great Lakes beach sand (Gallagher, 2007). Electrocompetent *E.coli* strain EC100 (Epicentre Technologies, Madison, WI) cells were used for electroporation of captured tet^R plasmids. EC100 cells are plasmid-free and are also tetracycline-sensitive.

Exogenous plasmid capture. Transmissible, tetracycline resistance (tet^R) plasmids were captured using a method developed in our laboratory (Figure 11) (Herrick *et al.*, 2014), modified from the original method of Fry and Day (1990). Cells were extracted from sediment samples to act as potential plasmid conjugation donors. Ten grams of sediment from each sediment sample were mixed with 90 mL of sterile 0.1% sodium pyrophosphate (pH 8) and agitated for 40 s. Sediment was allowed to settle at room temperature for 5 min. One milliliter of liquid was centrifuged for 10 min at 5,800 x g. The supernatant was removed and the remaining pellet was reconstituted with 1 mL of phosphate buffered saline (PBS) (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.4) and centrifuged for 10 min at 5,800 x g. The supernatant was removed and the pellet was reconstituted with 500 µL of PBS.

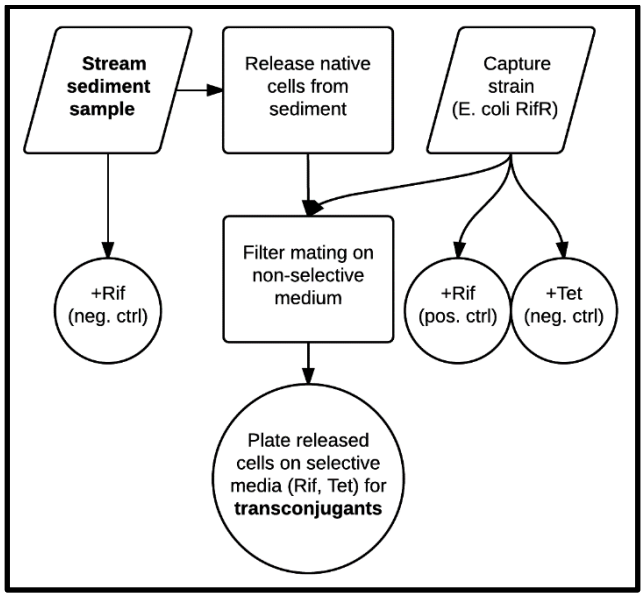

Figure 11. Exogenous Plasmid Capture method for capturing tetracycline resistance plasmids from sediment. Plasmids were captured from stream sediment samples by releasing cells from sediment and conjugating with a rifampicin-resistant strain of *E. coli*. Donors and recipients were individually plated onto TSA plates containing either rifampicin (50 µg/mL) or tetracycline (25 µg/mL) respectively, as negative controls (and to account for spontaneous rifampicin resistance), and onto TSA as a positive control. Transconjugants were selected on tetracycline-and rifampicin-amended medium.

Recipient strain LA61 (tet^S, rif^R), was cultured overnight in TSB at 37°C with agitation (~180 RPM). One milliliter of the culture was centrifuged at 5,800 x g for 10 min and washed with PBS in the same manner as the donor cells from the sediment. The resulting washed pellet was reconstituted in 500 µL of PBS.

The 500 µL volume of recipient cells was added to the centrifuge tube containing the 500 µL volume of potential donor cells. Cells were gently mixed by inversion for 20 s. Two hundred microliters of donor/recipient cell mixture were transferred onto a sterile 0.45 µm cellulose acetate filter (Millipore Corporation, Billerica, MA) applied to a TSA plate. The cells were allowed to conjugate for 24 hrs at 37°C. After incubation, the filter was aseptically removed from the TSA plate and placed in a 50 mL Falcon® tube. Ten

milliliters of PBS were added to the tube with the filter and vigorously vortexed for 3 min to release cells from the filter.

Five hundred microliters of the supernatant were pipetted onto TSA plates amended with rifampicin (50 μg/mL) and tetracycline (25 μg/mL) to select for transconjugants. Donors and recipients were individually plated onto TSA plates containing either rifampicin (50 μg/mL) or tetracycline (25 μg/mL) respectively, as negative controls (and to account for spontaneous rifampicin resistance), and onto TSA as a positive control. Putative transconjugant colonies were transferred to TSA plates containing tetracycline (25 μg/mL) and incubated for 24 hrs at 37°C to verify tetracycline resistance. Putative transconjugants were confirmed as LA61 using the BOX-PCR repetitive sequence fingerprinting method used as described in Gehr (2013) and Rademaker *et al.* (1998).

Plasmids were extracted from transconjugants using the plasmid preparation protocol described in Chapter 1 Methods (p. 18) and electroporated into a control recipient, EC100, to isolate single plasmids and to verify plasmid-borne tetracycline resistance.

Electroporation. Plasmid DNA was electroporated into electrocompetent tet^S *E. coli* strain EC100 (Epicenter Technologies, Madison, WI) using a 1 mm cuvette with a Bio-Rad GenePulser Xcell electroporator according to the manufacturer's instructions (Bio-Rad, Hercules, CA). EC100 cells were maintained and prepared for electroporation according to protocol 6.1.1 of the Bio-Rad Gene pulser instruction manual. Transformed cells were plated on TSA plates amended with 25 μg/mL tetracycline to verify the transformation of the cells with a tet^R plasmid.

Antibiotic susceptibility phenotyping. Modified Stokes assays were performed as described by Herrick *et al.* (2014). Resistance phenotypes for the antibiotics tetracycline (30µg), gentamicin (10 µg), ciprofloxacin (5 µg), trimethoprim/sulfamethoxazole (23.75 / 1.25 µg), imipenem (10 µg), tobramycin (10µg), kanamycin (30 µg), aztreonam (30 µg), ticarcillin (75 µg), piperacillin/tazobactam (100 / 10 µg), piperacillin (100 µg), and cefepime (30 µg) were assessed using a modified Stokes disk diffusion assay (Figure 12; Herrick *et al.*, 2014). Recipient EC100 control and tet^R transformants resulting from electroporation were cultured overnight in TSB. Sterile swabs were used to apply turbid broth cultures to Mueller-Hinton agar plates (Becton Dickinson). Electroporated plasmid-harboring cells were swabbed on the outer sections of the plate and the plasmid-free recipient was swabbed in the center third. Antibiotic diffusion disks were applied along the lines where the recipients and electrotransformants meet. Plates were incubated for 24 hours at 37°C.

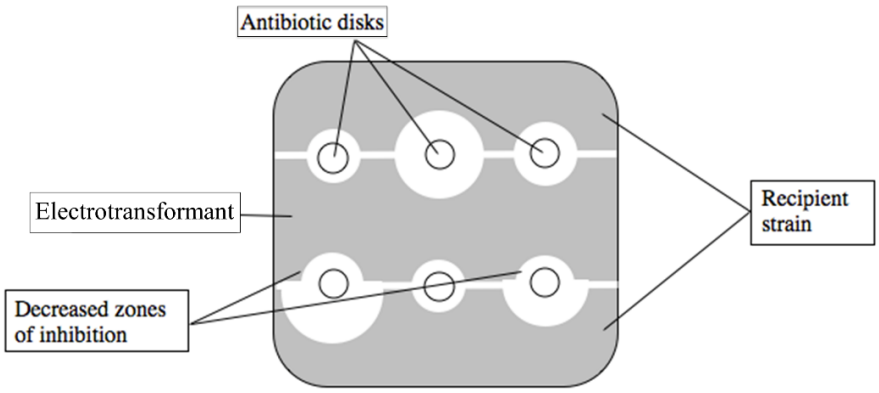

Figure 12. Modified Stokes Assay method for detecting decreased susceptibility conferred by electroporated plasmids. Transformants (electrocompetent *E.coli* EC100) strain and empty recipients (control) were applied to Mueller-Hinton media. Antibiotic diffusion disks were then placed along the lines at which the transformants and recipients meet. Resulting zones of inhibition of the transformants and the recipient were compared to determine if the plasmid in the transformant encodes resistance to the antibiotic in the diffusion disk. A reduction of ≥ 3 mm in the radius of the zone of inhibition was designated as "resistant" to the antibiotic.

The radii of clearing zones from the control side and the transformant side of the disks were measured using a caliper and compared. A ≥ 3 mm reduction in the radium of the zone of inhibition was considered an indication of resistance to the antibiotic (Acar and Goldstein, 1996).

***In vitro* restriction digests.** EcoRI *in vitro* restriction digests were performed using the methods described in Chapter 1 (p. 20). HindIII and SmaI *in vitro* restriction digests were performed by mixing 17.3 µL of plasmid DNA with 2 µL of buffer (Buffer J for SmaI, or buffer B for HindIII) (Promega, Madison, WI), 0.2 µL BSA (Promega), and 0.5 µL of either HindIII or SmaI restriction enzyme (Promega). Tubes were vortexed, briefly centrifuged, and incubated at 37°C (HindIII) or room temperature (SmaI) for 1 hr. Enzymes were inactivated at 70°C for 30 min. An agarose gel was prepared and loaded using the methods described in Chapter 1 to visualize the fragments and fragment sizes were estimated using a lambda/HindIII digest as a size standard (Chapter 1 Methods, p. 19).

Plasmid DNA sequencing. Plasmid DNA was extracted, prepared, and sequenced on the Illumina MiniSeq and the Oxford Nanopore MinION sequencers. For Illumina sequencing, plasmids were extracted from electrotransformants using the normal plasmid extraction protocol and were prepared for sequencing using Plasmid-Safe, EcoRI, and Nextera XT DNA Library Prep Kit (Chapter 1, pp. 18-25).

For ONT sequencing, plasmids were extracted from electrotransformants using the scaled-up plasmid extraction protocol and were prepared for sequencing using Plasmid-Safe, Sau3AI, and 1D Ligation Sequencing Kit (Chapter 1, pp. 2-30).

Sequence data processing. Sequence data was analyzed through a custom bioinformatics analysis pipeline (Figure 13, Table 5).

Table 5. Analysis software used.

Name of Software	Version Used	Authors and/or Reference	Link
Local Run Manager and GenerateFASTQ	1.3.1 and 1.0.0	Illumina	support.illumina.com/sequencing/sequencing_software/local-run-manager-for-miniseq/downloads.html
FastQC	0.11.7	Andrews, 2010	bioinformatics.babraham.ac.uk/projects/download.html#fastqc
MultiQC	1.2	Ewels *et al.*, 2016	multiqc.info/
NanoPlot	1.13.0	De Coster *et al.*, 2018	github.com/wdecoster/NanoPlot
Albacore	2.10.0	Oxford Nanopore Technologies	community.nanoporetech.com/downloads
Porechop	0.2.3	Wick *et al.*, 2017	github.com/rrwick/porechop
Filtlong	0.2.0	Ryan Wick	github.com/rrwick/filtlong
Unicycler	0.4.4	Wick *et al.*, 2017	github.com/rrwick/unicycler
QUAST	4.6.3	Mikheenko *et al.*, 2016	quast.sourceforge.net/quast
Bandage	0.8.1	Wick *et al.*, 2015	github.com/rrwick/bandage
Geneious	R8.1	Kearse *et al.*, 2012	www.geneious.com/download/
BWA	0.7.10	Li, 2013	bio-bwa.sourceforge.net/
Minimap2	2.10	Li, 2018	github.com/lh3/minimap2
Samtools	1.7	Li *et al.*, 2009	samtools.sourceforge.net/
Tablet	1.17.08.17	Milne *et al.*, 2017	ics.hutton.ac.uk/tablet/download-tablet/
Prokka	1.13	Seemann, 2014	github.com/tseemann/prokka
ABRicate	0.8	Torsten Seemann	github.com/tseemann/abricate
BLAST	2.8.0	Altschul *et al.*, 1990	blast.ncbi.nlm.nih.gov/Blast.cgi

Table 5. (cont.)

PlasmidFinder	1.3	Carattoli *et al.*, 2014	cge.cbs.dtu.dk/services/PlasmidFinder/
Integron_Finder	1.5.1	Cury *et al.*, 2016	github.com/gem-pasteur/Integron_Finder
EasyFig	2.2.2	Sullivan *et al.*, 2011	mjsull.github.io/Easyfig/

To generate fastq files, Illumina reads were basecalled, de-multiplexed, and adapters and barcode sequences were removed following the sequencing run. These processes were carried out automatically by the Local Run Manager v. 1.3.1 software on the Illumina MiniSeq sequencer (Illumina, San Diego, CA), using the analysis module Generate FASTQ v. 1.0.0.

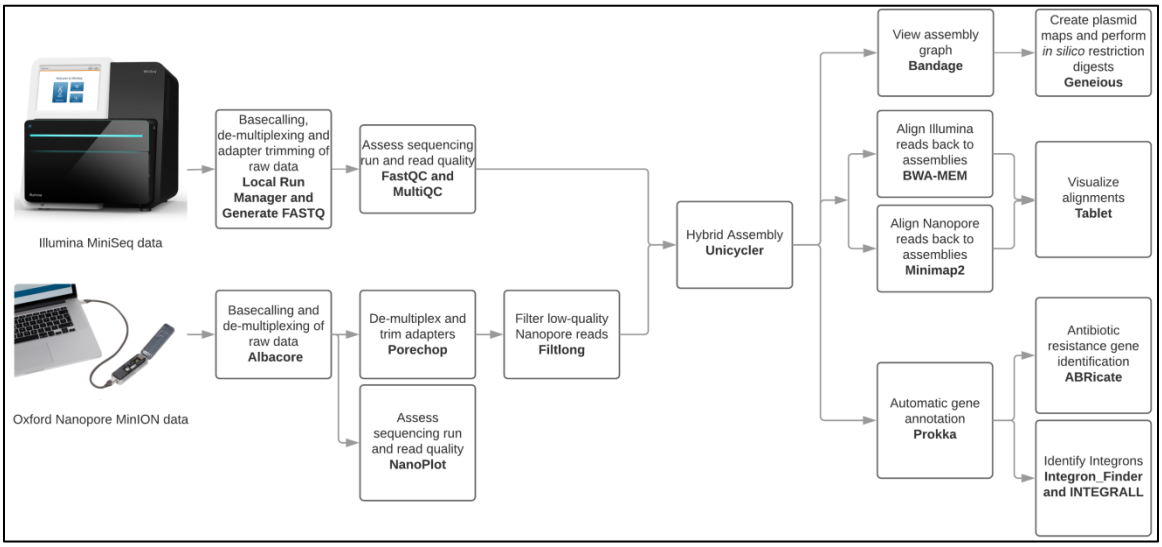

Figure 13. Bioinformatics analysis pipeline. MiniSeq image provided courtesy of Illumina, LLC and MinION image provided courtesy of Oxford Nanopore Technologies.

ONT reads were basecalled and de-multiplexed using Albacore v. 2.10.0 (Oxford Nanopore Technologies official basecalling software,

https://community.nanoporetech.com/downloads). However, the threshold for identifying barcode sequences within the reads is low (60% identity) and the probability of barcode misidentification is high. To prevent barcode misidentification, Porechop v. 0.2.3 (Wick *et al.*, 2017) was used to de-multiplex the reads using a stricter threshold (75% identity) and also to trim the barcode adapter sequences from the reads. Due to the relatively high error rate of ONT data, 10% of the ONT reads with the lowest quality scores were filtered using the Illumina reads as a reference (k-mer based quality filtering) using Filtlong v. 0.2.0 (https://github.com/rrwick/filtlong).

The quality of the Illumina reads was assessed using FastQC v. 0.11.7 as well as MultiQC v. 1.2 for generating quality score distribution plots (Andrews, 2010; Ewels *et al.*, 2016). NanoPlot v. 1.13.0 was used for assessing general sequencing run statistics and for creating figures displaying the quality and read length distributions of the ONT reads (De Coster *et al*., 2018).

Despite our attempt to remove the expected *E. coli* chromosomal DNA from the plasmid DNA samples prior to sequencing, some chromosomal DNA was sequenced along with the plasmids. BWA and Samtools (Li, 2013; Li *et al.*, 2009) were used to align the reads to the reference genome of the EC100 cells (*E. coli* K-12 DH10β, complete genome, accession: CP000948.1) with the intent to use only the non-aligning reads for assembly. This approach did not aid in assembling the plasmid genomes, as we discovered shared regions of sequence between the E. coli chromosome and two of the plasmids, pCCP1 and pCCP2. Thus we opted not to filter the reads prior to assembly.

Plasmid genome assembly. Plasmid genomes were assembled *de novo* using Unicycler v. 0.4.4 (Wick *et al.*, 2017). Unicycler was chosen for assembly because it

accepts both Illumina and ONT reads and was designed for assembling bacterial genomes and plasmids in a hybrid manner - taking advantage of the accuracy of the short reads and the length of the (less accurate) long reads (Wick *et al.*, 2017). The de-multiplexed Illumina reads and the de-multiplexed, filtered ONT reads were used to generate hybrid assemblies (Figure 13). The "--contamination" flag in Unicycler was used to supply the *E. coli* K12 DH10β reference genome in order to filter out any contaminating ONT reads. When the contamination flag is used, Unicycler will align ONT reads against the supplied reference genome, and will discard any reads that align to the reference better than the intermediate (short-read-only) assembly produced by Unicycler. Assembly graphs (gfa files) were visualized in Bandage v. 0.8.1 (Wick *et al.*, 2015) and evaluated using QUAST v. 4.6.3 (Mikheenko *et al.*, 2016).

Following the assembly, contigs were aligned to the *E. coli* K12 DH10β reference genome (accession: CP000948.1, minimum alignment length 200bp, e-value < $1e^{-6}$) using the BLAST search function that is integrated into the Bandage application. Any contigs that aligned to the reference, or were smaller than 1000 bp were removed, leaving only the circular contigs that were presumed to be the assembled plasmid genomes.

***In silico* and *in vitro* restriction digests.** Restriction digests were performed on each of the plasmids *in vitro* and *in silico*. The following restriction enzymes were used: EcoRI (restriction site - GAATTC), HindIII (restriction site - AAGCTT), and SmaI (restriction site - CCCGGG) and the DNA fragment patterns (number and molecular weight of fragments) seen on a gel were compared with those generated *in silico*.

In silico restriction digests were performed in Geneious v. R8.1 (Biomatters, Auckland, New Zealand) by uploading the assembled genome (in fasta format) to the

software (Kearse *et al.*, 2012). Restriction sites for all three enzymes mentioned above were identified within the plasmid genomes and the number and size (in bp) of the predicted fragments for each digest were calculated.

Read alignment. To determine the average depth of coverage, reads were aligned to their respective assemblies using BWA v. 0.7.10 (BWA-MEM, specifically) for the Illumina reads and Minimap2 v. 2.10 for the ONT reads (Li, 2013; Li, 2018). These two aligners were chosen because BWA-MEM was originally designed for highly accurate short reads ranging from 75-300 bp while Minimap2 was designed for error-prone long reads (greater than 300 bp), produced by either ONT or PacBio single molecule sequencers.

After alignment files (SAM format) were generated with either BWA or Minimap2, SAM files were converted to BAM files using SAMtools v. 1.7 with the SAMtools command "view" (Li *et al.*, 2009). SAMtools command "sort" was used to sort the BAM files, specifically sorting the aligned reads within the BAM file based on their position in reference genome (sorted from first to last base in the assembly). SAMtools command "flagstat" was used to generate statistics on alignments and SAMtools command "index" was used to generate index files that are necessary for viewing alignments in Tablet. Tablet v. 1.17.08.17 was used to visualize the alignment BAM files (Milne *et al.*, 2017).

Annotation. The plasmid genome assemblies were annotated automatically using Prokka v. 1.13 (Seemann, 2014) and then curated manually. ABRicate v. 0.8 (Seemann, github.com/tseemmann/abricate/) was used with the ARG-ANNOT database of antibiotic resistance genes to screen the assemblies for antibiotic resistance genes not identified by

Prokka. PlasmidFinder v. 1.3 (Carattoli *et al.*, 2014) was used to identify replication sequences present in the assemblies to identify the incompatibility group that each plasmid belongs to. Integron_Finder v. 1.5.1 was used to identify integrons as well as CALIN elements (Cluster of attC sites Lacking Integrase Nearby) (Cury *et al.*, 2016). The INTEGRALL database, which contains over 9000 entries for a variety of integrons, integrases, and gene cassettes, was queried to confirm Integron_Finder's results and to identify additional integron-related sequences for each of the plasmids (Moura *et al.*, 2009). Lastly, the amino acid sequences of the remaining coding sequences (also CDS) identified by Prokka as "hypothetical protein" were aligned to the non-redundant (nr) GenBank database using BLASTp (Altschul *et al.*, 1990). CDS were annotated when an alignment occurred to a known protein with a minimum % identity of 87%, >92% query coverage, and e-value < $1e^{-6}$.

Easyfig v. 2.2.2 was used to perform pairwise alignments and compare parts or whole genomes to one another (Sullivan *et al.*, 2011). Geneious v. R8.1 was used to create plasmid genome maps (Kearse *et al.*, 2012).

Results and Discussion

Exogenous plasmid capture. Transmissible tet^R plasmids were isolated from stream sediment using an exogenous plasmid capture method that does not require the cultivation of donor cells. Cells were released from stream sediment samples from each of the sampling sites and combined with the rif^R recipient strain, LA61. During each capture attempt, positive and negative controls were used to verify conjugation.

Eleven captures were performed from February 2016 to April 2017 and four of the eleven captures yielded a total of 77 transconjugants exhibiting resistance to tetracycline (Table 6). No transconjugants were yielded from either Muddy Creek, War Branch, or Pleasant Run sites in all of the captures that were performed.

Table 6. Summary of exogenous plasmid capture results.

Date of capture	Number of transconjugant colonies	Site
July 2016	2	Cooks Creek Park
September 2016	2	Cooks Creek Park
October 2016	6	Cooks Creek Rt. 11
February 2017	77	69 from Cooks Creek Rt. 704 8 from Cooks Creek Park

Three plasmids were selected for further investigation: pCCP1 and pCCP2 which were both captured in July 2016 from Cooks Creek Park, and pCCRT11-6 which was captured in October 2016 from Cooks Creek at Route 11. In addition, one plasmid conferring multi-drug antibiotic resistance, pEG1-06, was investigated in this study. It was captured previously in October of 2006 from Shull's Run in Rockingham County,

Virginia (Gehr, 2013) using the same exogenous plasmid capture protocol and using the same recipient strain, LA61.

Verification of LA61 transconjugants. Putative transconjugants were verified as LA61 (harboring a tetracycline resistance plasmid) through plating on media amended with both rifampicin and tetracycline. Transconjugants were also cultivated on Eosin Methylene Blue plates and after 24 hrs incubation, dark purple colonies with a metallic green sheen were observed – typical of *E. coli* - which was also observed with the LA61 plasmid-free recipient. A PCR-based DNA fingerprinting method was used to confirm transconjugants were LA61. The BOX rep-PCR amplifies intervening regions from 154 bp BOX sequences that are interspersed throughout many prokaryotic genomes in strain-specific locations (Rademaker *et al.*, 1998). Each transconjugant and LA61 were also subjected to BOX rep-PCR and the banding patterns were identical (Figure 14).

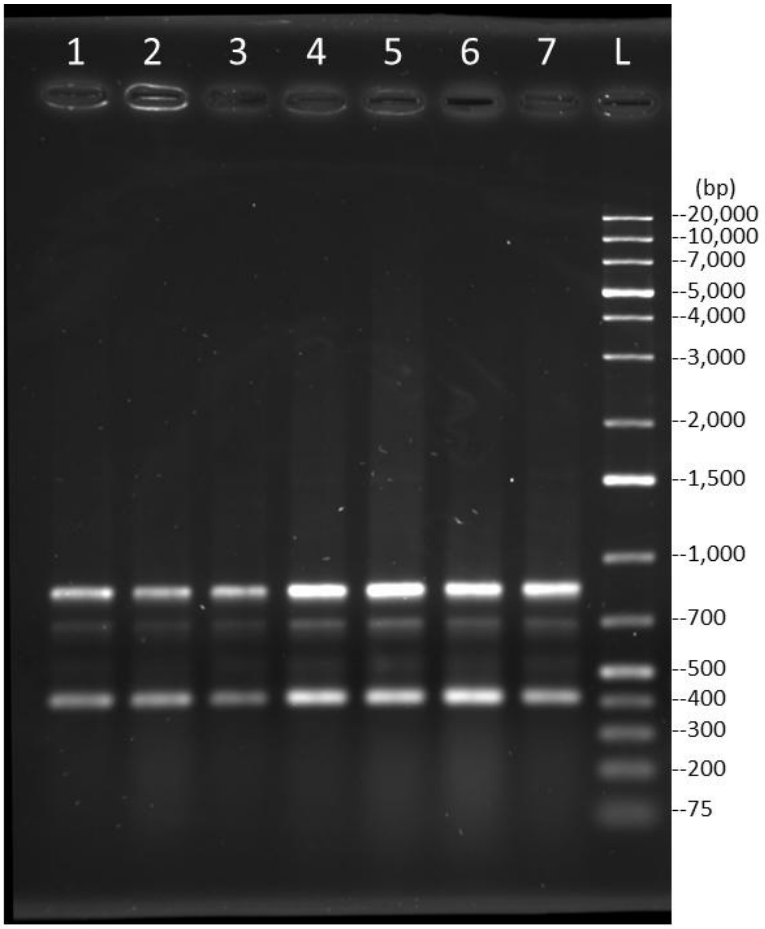

Figure 14. Example of BOX rep-PCR fingerprint analysis of transconjugants. Lane 1: LA61, lane 2: transconjugant from Cooks Creek Rt. 704, lane 3: transconjugant from Cooks Creek park, lane 4: LA61, lane 5: LA61, lane 6: transconjugant from Cooks Creek Rt. 704, lane 7: transconjugant from Cooks Creek Rt. 704, lane L: GeneRuler 1 kb Plus DNA ladder (ThermoFisher).

Antibiotic susceptibility phenotyping. Plasmids were extracted from transconjugants and electroporated into the plasmid-free, tetracycline sensitive, and electrocompetent *E. coli* strain EC100. Tetracycline resistance of electrotransformants was verified by plating on a TSA plate amended with tetracycline. These were subsequently tested for resistance to 11 additional antibiotics using a modified Stokes antibiotic disc diffusion assay (Table 7, Figure 12). pEG1-06 conferred resistance to tetracycline, ticarcillin, piperacillin, and piperacillin/tazobactam. The pEG1-06 plasmid

also conferred decreased susceptibility to cefepime; however the resistance breakpoint of 3 mm reduction of the zone of inhibition relative to the plasmid-free EC100 strain was not surpassed (exhibited a 2 mm reduction). Plasmids pCCP1, pCCP2, and pCCRT11-6 each conferred only tetracycline resistance.

Table 7. Antibiotic resistance phenotypes of four tet^R plasmids, electrotransformed into *E. coli* strain EC100.

Plasmid	Resistance Phenotype[a]
pEG1-06	TE, TIC, PIP, TZP, FEP[b]
pCCP1	TE
pCCP2	TE
pCCRT11-6	TE

[a] Tetracycline (TE), ticarcillin (TIC), piperacillin/ tazobactam (TZP), piperacillin (PIP), and cefepime (FEP).
[b] Susceptibility was decreased, but did not surpass the resistance breakpoint of a reduction in 3 mm in the zone of inhibition

Plasmid DNA sequencing. Plasmid DNA was extracted, prepared, and sequenced on the Illumina MiniSeq and the Oxford Nanopore MinION sequencers. For Illumina sequencing, plasmids were extracted from electrotransformants using the normal plasmid extraction protocol and were prepared for sequencing using Plasmid-Safe, EcoRI, and Nextera XT DNA Library Prep Kit (Chapter 1 Methods pp. 5-12). For ONT sequencing, plasmids were extracted from electrotransformants using the scaled-up plasmid extraction protocol and were prepared for sequencing using Plasmid-Safe, Sau3AI, and 1D Ligation Sequencing Kit (Chapter 1 Methods pp. 2-17).

Plasmid genome assembly. After de-multiplexing and barcode trimming, the ONT reads were filtered using Filtlong to remove the 10% of reads that had the lowest quality scores. These quality scores are arbitrary k-mer matching quality scores set by Filtlong and are not to be confused with Phred Q scores.

Unicycler is a *de novo* hybrid assembly pipeline designed for bacterial genomes that assembles short reads into an accurate and connected De Bruijn assembly graph using SPAdes followed by simplification of the graph using both long and short reads (Wick *et al.*, 2017). Unicycler was first used to assemble the plasmid genomes using the short reads only (SRO). This resulted in assemblies with a large number of contigs for each plasmid, ranging from 11 contigs for pEG1-06 up to 947 contigs for pCCP1 with varying degrees of connectivity (Table 8, Figure 15A, 14D, 14G, 14J). In addition to the expected contaminating *E. coli* chromosomal contigs present in each of the assemblies, we expected each assembly to result in one contig (connected with no dead ends) of much greater sequencing depth that we presumed would belong to the plasmid genome. In the SRO assemblies for each plasmid, contigs connected into roughly circular structures that had no dead ends (at a higher depth); however they consisted of three or more inter-connected contigs, and could not be resolved into a single circular contig, likely due to repeated sequences that were larger than the length of the Illumina reads (max 151 bp reads length) (Figure 15A, 14D, 14G, 14J). Difficulty in resolving plasmid structures from whole-genome sequence data, especially large (> 50kb) plasmids with repeated sequences, is a long-standing issue that can theoretically be solved with the addition of long sequence reads, even those with low (< 10X) depth of coverage (Arredondo-Alonso *et al.*, 2017; Wick *et al*., 2017).

Unicycler was used to assemble the plasmid genomes using both the Illumina reads and the ONT reads. The hybrid assembly of each plasmid had fewer, larger contigs with less dead ends than the SRO assemblies, however many small (<1000 bp), linear contigs with dead ends were present (Table 8, Figure 15B, 14E, 14H, 14K). These small,

linear contigs likely were assembled from the contaminating *E. coli* reads or were assembly errors arising from the plasmid reads.

Following assembly of the plasmid genomes, contiguous sequences (contigs) smaller than 1000 bp and contigs that aligned to the *E. coli* reference genome (*E. coli* str. K12 substr. DH10β [same strain as EC100], accession: CP000948.1) were removed. Assemblies were evaluated using QUAST (Table 8) and assembly graphs were visualized using Bandage (Figure 15) (Mikheenko *et al.*, 2016; Wick *et al.*, 2015).

Table 8. Assembly metrics for four plasmids assembled using only short reads (SRO) and short plus long reads (Hybrid). QUAST was used to determine these assembly statistics.

Plasmid	Assembly type	# of contigs	# of contigs ≥ 1000 bp	Largest contig (bp)	Total length of assembly (bp)	%GC	N50[a] (bp)
pCCP1	SRO	947	891	78,847	3,938,113	51.16%	6,060
pCCP1	hybrid	1	1	59,842	59,842	57.04%	59,842
pCCP2	SRO	217	181	140,875	4,415,838	50.83%	44,771
pCCP2	Hybrid	1	1	59,842	59,842	57.09%	59,842
pCCRT11-6	SRO	894	736	106,131	1,535,433	51.79%	1,724
pCCRT11-6	Hybrid	1	1	121,469	121,469	48.89%	121,469
pEG1-06	SRO	11	9	60,245	84,141	60.96%	60,245
pEG1-06	Hybrid	1	1	71,416	71,416	62.74%	71,416

[A] N50 is the contig length (in bp) that 50% of the assembly is contained in contigs that are greater than or equal to the length of N50

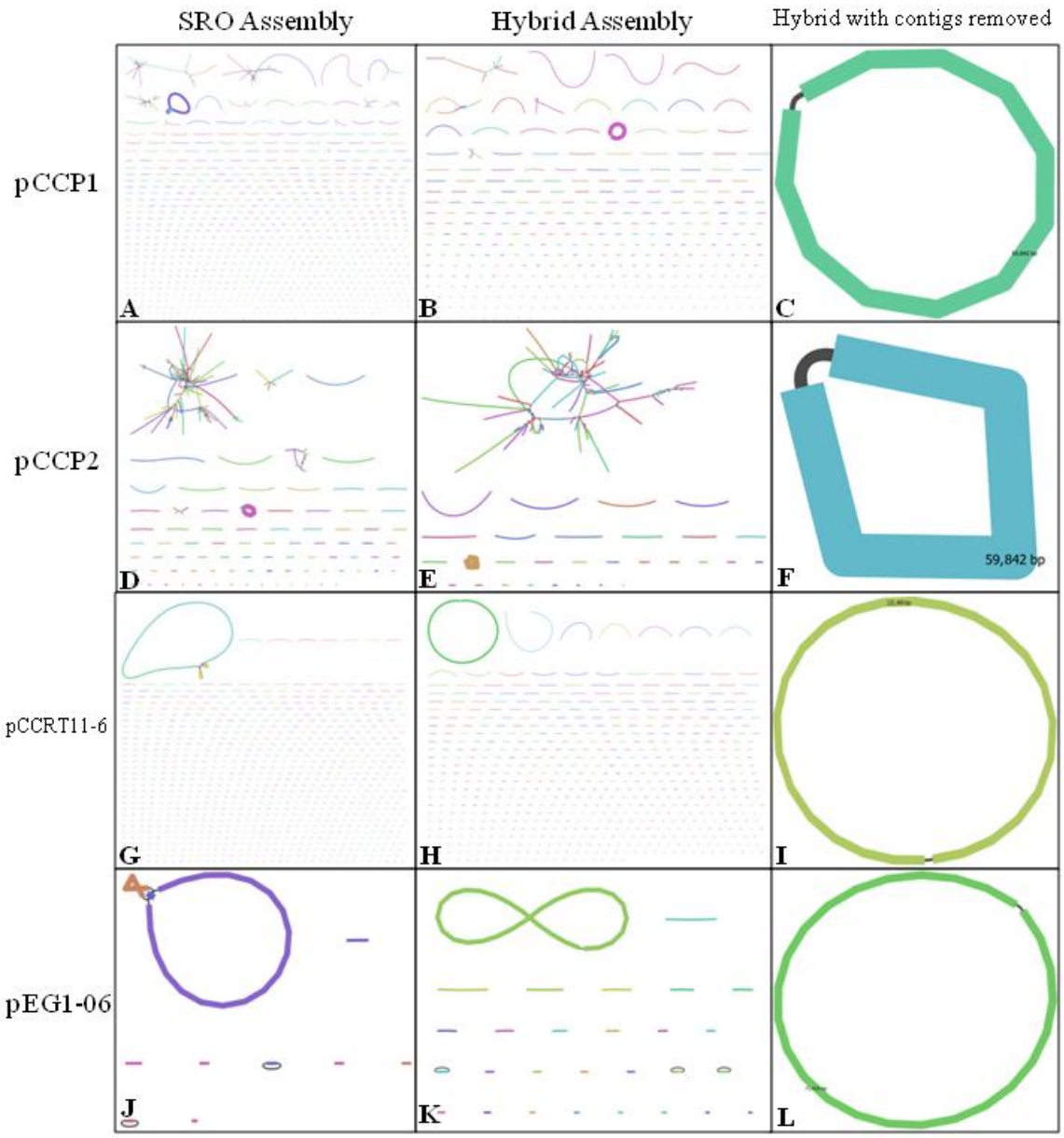

Figure 15. Short read only assembly graphs (SRO), hybrid assembly graphs, and hybrid assembly graphs after contigs < 1000 bp or those that aligned to *E. coli* were removed for plasmids pCCP1 (A, B, C), pCCP2 (D, E, F), pCCRT11-6 (G, H, I), pEG1-06 (J, K, L). Node (contig) colors are random and between different assembly graphs, contig sizes are not drawn to scale. Assembly graph files (gfa) were generated using Unicycler (Wick *et al.*, 2017) and visualized using Bandage (Wick *et al.*, 2015).

Hybrid assembly of pCCRT11-6 after removing contigs < 1000 bp or those that aligned to *E. coli* resulted in a single, circular contig that was 121,469 bp in length and had 48.89% GC content (Figure 15I, Table 8). Assembly of pCCP1 resulted in a single,

circular contig that was 59,842 bp in length and had 57.04% GC content (Figure 15C, Table 8). Assembly of pCCP2 resulted in a single, circular contig that was 59,842 bp in length and had 57.09% GC content (Figure 15F, Table 8). Assembly of pEG1-06 resulted in a single, circular contig that was 71,416 bp in length and had 62.74% GC content (Figure 15L, Table 8).

***In silico* and *in vitro* restriction digests.** Assembly errors oftentimes arise in plasmid genomes due to large duplications or repeats because these regions cannot be resolved by short read sequencing alone (Smalla *et al.*, 2015). To help verify that the plasmid genomes were correctly assembled, both *in vitro* and *in silico* restriction digests were performed on each of the plasmids. Each plasmid was digested *in vitro* with EcoRI, SmaI, and HindIII and visualized on a gel. *In silico* digests were performed using Geneious software to determine the predicted number of fragments and their respective sizes in bp. The number and size of the resulting plasmid DNA fragments were compared between both types of "digest".

For pCCP1 digested by EcoRI, 15 out of 16 *in silico* predicted fragments were visible on a gel (Figure 16). The smallest fragment, predicted to be 369 bp in length, was not observed. For pCCP1 digested by HindIII, 6 out of 6 *in silico* predicted fragments were observed and for pCCP1 digested by SmaI, 13 out of 16 predicted fragments were observed (Figure S19). The 3 fragments not observed were predicted to be 766, 40, and 40 bp in size.

pCCP1 *in silico* EcoRI digest	
Fragment Size (bp)	**Seen on gel image?**
14,139	Yes
9,576	Yes, bands overlap
9,056	Yes, bands overlap
6,256	Yes
4,952	Yes
4,293	Yes, bands overlap
4,240	Yes, bands overlap
2,378	Yes
1,344	Yes
1,089	Yes
845	Yes
686	Yes
619	Yes
369	No

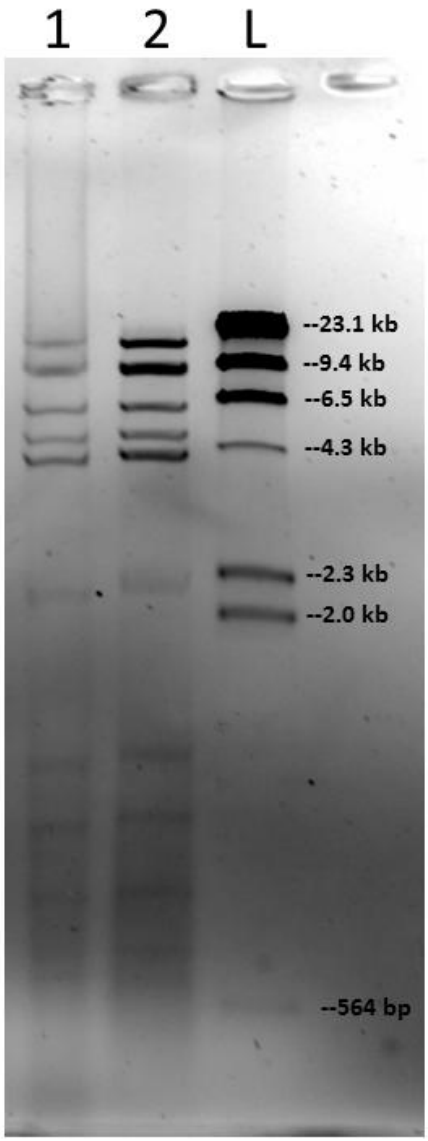

Figure 16. Example of *in silico* and *in vitro* restriction digestion of pCCP1 with EcoRI. *In silico* digest fragment number and sizes were determined using Geneious. Lane 1: pCCP1 treated with RNAse A, plasmid-safe, and EcoRI, lane 2: same sample as lane 1 with AMPure XP bead purification and concentration, lane L: lambda/HindIII digest.

For pCCP2 digested by EcoRI, 9 out of 14 fragments were observed and the fragments not observed were predicted to be 1,344, 845, 686, 619, and 319 bp in length (Figure S19). For pCCP2 digested by HindIII, 6 out of 6 fragments were observed and for pCCP2 digested by SmaI, 13 out of 16 predicted fragments were observed (Figure S19). The 3 fragments not observed were predicted to be 766, 40, and 40 bp in size.

For pCCRT11-6 digested by EcoRI, 6 out of 8 fragments were observed and the fragments not observed were predicted to be 600 and 284 bp in length (Figure S16). For pCCRT11-6 digested by HindIII, 1 fragment was predicted and observed (Figure S17). For pCCRT11-6 digested by SmaI, 2 out of 3 fragments were observed and the fragment not observed was predicted to be 766 bp in length (Figure S18).

For pEG1-06 digested by EcoRI, 6 out of 6 fragments were observed (Figure S13) and for pEG1-06 digested by HindIII, 5 out of 9 predicted fragments were observed (figure S14). The 4 fragments not observed were predicted to be 717, 159, 159, and 59 bp in size. For pEG1-06 digested by SmaI, 3 out of 3 predicted fragments were observed (Figure S15).

All of the predicted fragments that were not observed within the agarose gels were ≤ 1,344 bp in size and were likely not seen for three reasons: (1) these fragments were too small (e.g. 40 bp), and ran off the gel due to the long gel running time (90 min or longer in some gels); (2) the DNA concentrations of these fragments were too low, thus were too faint to be seen on the gel; or (3) these fragments did appear on the gel, but were masked by a light colored patch near the 564 bp marker that was caused by the loading dye that was mixed with all samples prior to loading the gels (Figure S19).

Read alignment to assembled genomes. Reads for each plasmid were aligned to their respective assemblies using BWA (Li, 2013) for the Illumina reads and Minimap2 (Li, 2018) for the ONT reads. Alignment statistics were determined using Samtools (Li *et al.*, 2009) and Tablet (Milne *et al.*, 2017). For all 4 plasmids, 100% of the assemblies were covered by both types of reads (Table 7). Visualizations of pEG1-06 Illumina reads

aligned to its assembly and of pCCP2 ONT reads aligned to its assembly can be found in the supplementary materials (Figure S11 and S12).

Table 9. Read alignment statistics of the Illumina and ONT reads aligned to each of the four plasmid genome assemblies. Average coverage depth, max coverage depth and % of assembly covered were determined using Tablet and % of reads aligned to assembly was determined using Samtools.

Illumina reads					
Plasmid	**Genome size**	**Average coverage depth (X)**	**Max coverage depth (X)**	**% of reads aligned to assembly**	**% of assembly covered**
pEG1-06	71,416 bp	4,373.1	8,333	98.95	100
pCCRT11-6	121,469 bp	1,847.1	4,730	93.85	100
pCCP1	58,942 bp	1,929.8	4,552	76.62	100
pCCP2	58,942 bp	1,626.5	5,107	50.2	100
ONT reads					
pEG1-06	71,416 bp	2,469.9	2,691	89.08	100
pCCRT11-6	121,469 bp	530.6	697	77.08	100
pCCP1	58,942 bp	105.8	178	24.8	100
pCCP2	58,942 bp	114.7	194	42.4	100

Average depths of coverage for both types of reads aligned to each of the assemblies were high. This was expected due to the relatively small size of the plasmid genomes and the small number of samples that were multiplexed and sequenced in each of the sequencing runs. Accurate *de novo* draft assemblies can typically be achieved with anywhere from 50 - 100X coverage of bacterial genomes (Desai *et al.*, 2013). 24.8% of ONT reads from pCCP1 aligned to the assembly, which was much lower than the three other plasmids and their respective ONT read alignments. This sample had a high percentage of reads belonging to the *E. coli* chromosome, as confirmed by BLAST alignments performed in bandage and the filtering performed after hybrid assembly (Figure 15B).

Plasmid genome annotation. Plasmid genome assemblies were annotated using Prokka, which is a tool designed for rapidly annotating bacterial genomes by predicting coding sequences, translating predicted CDS into amino acid sequences, and comparing the amino acid sequences to databases of known proteins using BLAST (Seemann, 2014). CDS that align to a protein with an e-value < 10^{-6} were annotated as the reference (i.e. the protein from the database that the CDS aligned to) and CDS with no alignments < 10^{-6} were annotated as "hypothetical protein." Then the plasmid genomes were screened for antibiotic resistance genes using ABRicate with the ARG-ANNOT database (Antibiotic Resistance Gene-ANNOTation) and screened for integrons and integron-related sequences using Integron_Finder and the INTEGRALL database. Lastly, the amino acid sequences of CDS that were labeled as "hypothetical protein" were aligned to the nr database using BLASTp. These data are summarized in Table 8.

Table 10. Summary of isolation sources, sizes, CDS, incompatibility groups, and resistance genes of the four tetracycline resistance plasmids.

Plasmid	Sediment Sample source	Size (bp)	# Annotated CDS	# CDS classified as "hypothetical protein"	Inc group	Resistance genes
pEG1-06	Shulls Run	71,416	74	7	P-1β	*tetG, tetC, tetR, strB, folP, pse-1, emrE, ant1, ydhC*
pCCRT11-6	Cooks Creek at Rt. 11	121,469	54	83	A/C_2	*tetA, tetR, bcr, strB, yedA*

Table 10. (cont.)

pCCP1	Cooks Creek Park	58,942	71	34	P-9	*tetR, tetA, yedA*
pCCP2	Cooks Creek Park	58,942	71	35	P-9	*tetR, tetA, yedA*

Plasmid pEG1-06. The 71,416 bp pEG1-06 genome is comprised of 81 CDS, of which 74 were successfully annotated, and 7 were designated as "hypothetical proteins" (Figure 17, Table S1). The replicon *trfA* (plasmid replication initiator protein) was identified by Prokka and confirmed by PlasmidFinder as an IncP-1β replicon that aligned with 99.14% identity and 100% coverage to IncP-1β plasmid R751 from an *Enterobacter aerogenes* isolate (accession U67194.4, Thorsted *et al.*, 1985). The pEG1-06 genome also shared a high degree of similarity with pEG1-1 (99.99% identity, 96.7% coverage), which was captured at the same time as pEG1-06 (Gehr, 2013), sequenced, and identified as an IncP-1β plasmid (Libuit, 2016).

The pEG1-06 genome includes 8 functional modules: replication, stability, accessory module 1 (AM1), replication, mate pair formation (Mpf), accessory module 2 (AM2), DNA transfer region (Dtr), and stability (Figure 17). pEG1-06 is likely a self-transmissible (i.e. conjugative) plasmid due to its initial conjugation during capture (Gehr, 2013) and the presence of both Mpf and the Dtr modules that are found in all IncP-1β plasmids (Adamczyk and Jagura-Burdzy, 2003). The ~14 kb Mpf module consists of the genes *trbB/C/D/E/F/G/H/I/J/K/L/M/N/O/P* and *ompX*; these genes code for

proteins that make up the type IV secretion system used during conjugation (Smillie *et al.*, 2010). The ~15 kb Dtr module consists of the genes *traC/D/E/F/G/I/J/K/L/M/N/O* which are known to be involved in relaxosome formation and the initiation of rolling circle replication necessary for conjugal DNA transfer (Adamczyk and Jagura-Burdzy, 2003).

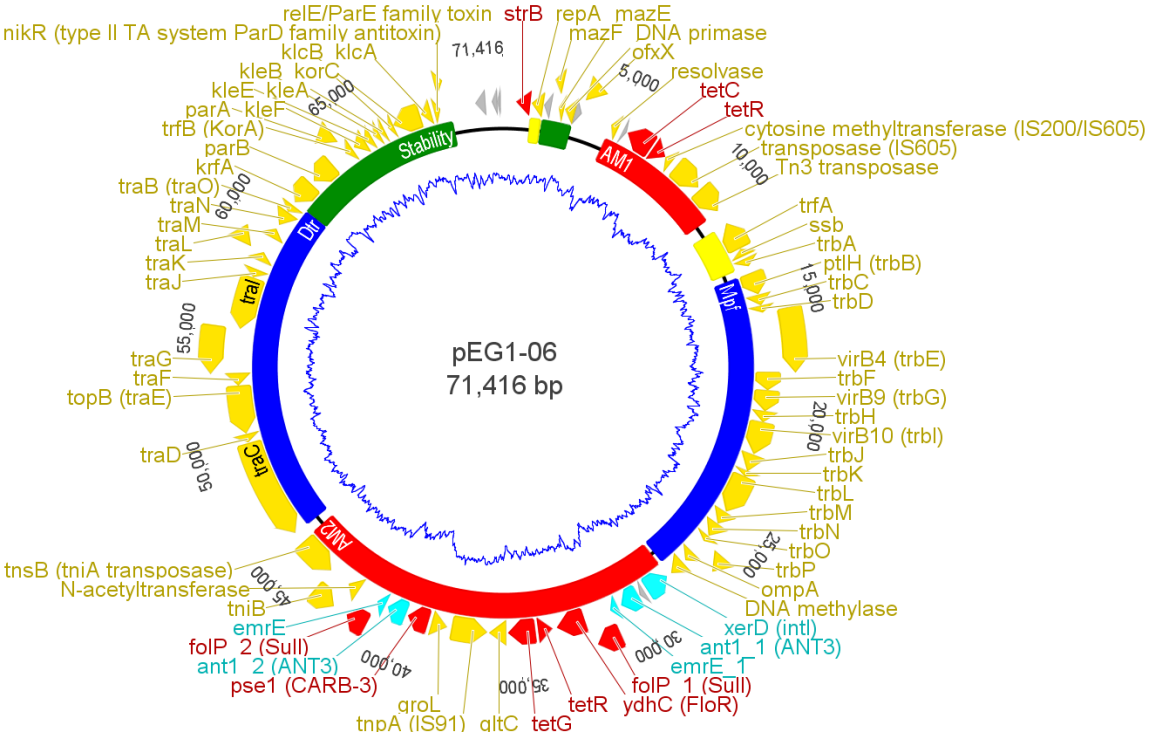

Figure 17. Functional modules and CDS mapped to the pEG1-06 genome assembly. Coding sequences and annotations are on the outer circle and the arrows point in the direction of transcription (outermost CDS are on complementary strand and the inner CDS are on the template strand): CDS annotated by Prokka and confirmed by BLASTp search (gold), antibiotic resistance genes annotated with ABRicate (red), Integron_finder annotations (cyan), and CDS annotated by Prokka as "hypothetical protein" (grey). Plasmid functional modules labeled on the inner circle: Accessory Modules (AM) (red), stability and partitioning regions (green), DNA transfer (Dtr) and mate pair formation (Mpf) regions (blue), and replication regions (yellow). Inner blue trace represents % GC.

Another defining feature of IncP-1β plasmids is that the Mpf and Dtr modules are often separated by clusters of restriction sites, allowing for the integration of accessory modules that are non-essential to the plasmid genome (Adamczyk and Jagura-Burdzy,

2003). Between the Dtr and Mpf modules, the pEG1-06 genome contains a complete class 1 integron identified by Integron_Finder (Figure 18). The ~17 kb integron consists of *intI* integrase (tyrosine recombinase) and its promotor Pint_1, gene cassette promotor Pc_1, integrase attachment site *attI*, a hypothetical protein, and two gene cassettes. The first cassette contains a recombination site *attC* and *ant1*, which encodes an aminoglycoside resistance protein, streptomycin 3"-adenylytransferase. The second cassette contains a recombination site *attC* and *emrE*, which encodes a multi-drug transporter protein. This class 1 integron contains the three key components which classify it as a complete integron: (1) an *intI* gene encoding for an integrase that controls the integration and excision of gene cassettes, (2) an *attI* site, and (3) a gene cassette promotor (Pc) that drives the expression of the cassettes (Schlüter *et al.*, 2007).

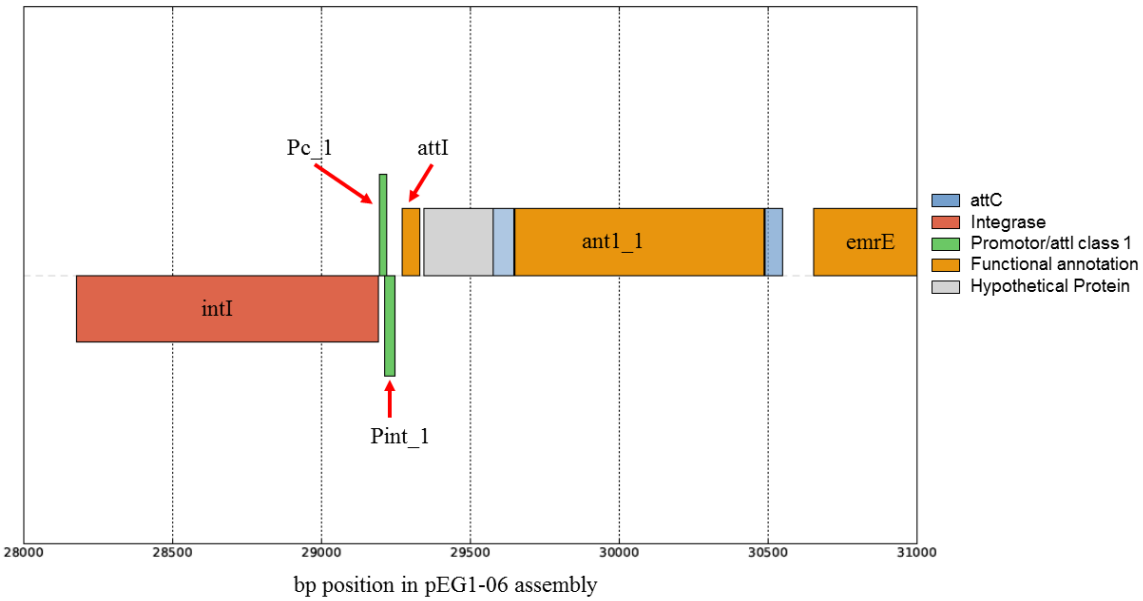

Figure 18. Class 1 integron located in accessory module 2 of the pEG1-06 genome. The integrase *intI* and its promotor are positioned below the other annotations to indicate transcriptional direction to the left, and the annotations above are positioned to indicate transcriptional direction to the right. Predicted CDS for *ant1_1*, aminoglycoside resistance gene, and *emrE*, multidrug transporter gene, were confirmed by Prokka.

A CALIN element (Cluster of *attC* sites Lacking Integrase Nearby) was also identified by Integron_Finder that is ~8 kb downstream of the class 1 integron, beginning with β-lactamase gene *pse-1* (914 bp), *attC* (110 bp), aminoglycoside resistance gene *ant1* (809 bp), attC (59 bp), and multidrug transporter gene *emrE* (347 bp). The % GC of the CALIN element and the class 1 integron have obvious derivations from the mean % GC of the entire pEG1-06 genome and suggests that these regions of the genome were acquired in recent evolutionary history by horizontal gene transfer (Figure 17).

In addition to the insertion site between the Mpf and Dtr modules another classic insertion site in IncP-1β plasmid genomes occurs between the origin of vegetative replication (*oriV* - likely located directly upstream of *repA* in pEG1-06) and *trfA*, which splits the set of replication genes into two parts (Dennis, 2005). IncP-1β plasmids are known to incorporate additional genes in these specific locations because they do not disrupt any of the essential backbone gene functions. Thus, insertions are common because they likely do not produce significant deleterious effects. The accessory module 1 present in pEG1-06 may have been incorporated by a transposon, due to the separation of *repA* and *trfA* plasmid replication genes and the presence of the transposon-related genes, Tn3 transposase, transposase (IS605), and a gene encoding a putative DNA resolvase (Figure 17).

pEG1-06 resistance phenotype and genotype. Plasmid pEG1-06 conferred resistance to tetracycline, piperacillin, ticarcillin, piperacillin/tazobactam and conferred decreased susceptibility to cefepime (Table 7). These phenotypes are likely due to the presence of multiple antibiotic resistance genes located throughout the genome including the tetracycline resistance genes *tetC*, *tetR* (repressor), *tetG* and the β-lactamase gene *pse-*

1 for resistance to piperacillin, ticarcillin, and piperacillin/tazobactam as well as *pse-1* for decreased susceptibility to cefepime.

Surprisingly, pEG1-06 did not confer resistance to trimethoprim/sulfamethoxazole despite the presence of 2 copies of sulfonamide resistance gene *folP* (also known as *sul1*), suggesting these genes may be non-functional. Additional antibiotic and toxin resistance genes located within the genome that were not tested phenotypically were the streptomycin/spectinomycin resistance gene *strB*, two copies of the streptomycin/spectinomycin resistance gene *ant1*, the florfenicol resistance gene *ydhC* (also known as *floR*), and the multi-drug transporter gene *emrE*, which is known to confer resistance to toxic compounds such as methyl viologen, ethidium bromide, acriflavine, tetraphenylphosphonium, and benzalkonium (Yerushalmi *et al.*, 1995).

pEG1-06 stability modules. The modules designated as "stability" (~8kb and 0.5 kb, respectively) contain essential backbone genes with predicted functions including the plasmid partitioning genes *parB/A;* the plasmid maintenance and stability genes *krfA*, *trfB, kleF/E/B/A, korC, klcB/A, nikR* (type II TA system *parD* family antitoxin); and the *relE/parE* family toxin gene. A second pair of toxin antitoxin (TA) genes were identified, the *mazF/E* Type II TA system. These are genes involved with post-segregational killing or so-called plasmid-addiction systems and are common to large, conjugative plasmids (Schlüter *et al.*, 2015). Toxin *mazF* encodes an endoribonuclease that targets RNAs produced in the cell, and *mazE* encodes an antitoxin that binds to and degrades the mazF toxin (Kamada *et al.*, 2003). If a daughter cell does not inherit a plasmid copy following cell division, it will not be able to produce the antitoxin protein, thus killing plasmid-free

daughter cells. The *nikR* (*parD*) and *parE* genes are known to be type II antitoxin and toxin genes respectively, where *parD* antitoxin either inhibits the toxin itself or inhibits the synthesis of the toxin (or potentially both) and the *parE* toxin inactivates DNA-gyrase, thus inhibiting DNA replication leading to cell death (Schlüter *et al.*, 2007).

An interesting characteristic that was observed with pEG1-06 was that when the plasmid DNA was extracted from EC100 cells, the cells containing pEG1-06 routinely yielded significantly more plasmid DNA than the three other plasmids investigated in this study. When pEG1-06 DNA was visualized in agarose gels, regardless whether undigested plasmid or digested plasmid samples were visualized on an agarose gel, pEG1-06 always produced the brightest bands and routinely yielded higher DNA concentrations when measured using the Qubit fluorimeter. We hypothesize that this may be caused by the presence of the two different types of TA systems present within the pEG1-06 genome in addition to the various genes related to stable inheritance (*kfrA*), active partitioning (*korA*, *parB/A*), and efficient modulation of partitioning (*kleF/E/B/A*, *klcB/A*) (Schluter *et al.*, 2007).

Plasmid pCCRT11-6. The 121,469 bp pCCRT11-6 genome is comprised of 137 CDS of which 54 were successfully annotated and 83 were designated as "hypothetical proteins" (Figure 19, Table S2). The CDS for the replicon *repA* was identified by PlasmidFinder as *repA* with the closest match from the IncA/C_2 plasmid pNDM-KN originally isolated from *Klebsiella pneumoniae* strain Kp7 (accession JN157804, Caratolli *et al.*, 2012). The *repA* genes aligned with 92.27% identity and 100% coverage. The entire pCCRT11-6 genome is most similar to the 173 kb, IncA/C_2, type 1 plasmid

R16a originally isolated from *Providencia stuartii* and aligned with 87% identity and 74% coverage of the pCCRT11-6 genome (Szabo *et al.*, 2016).

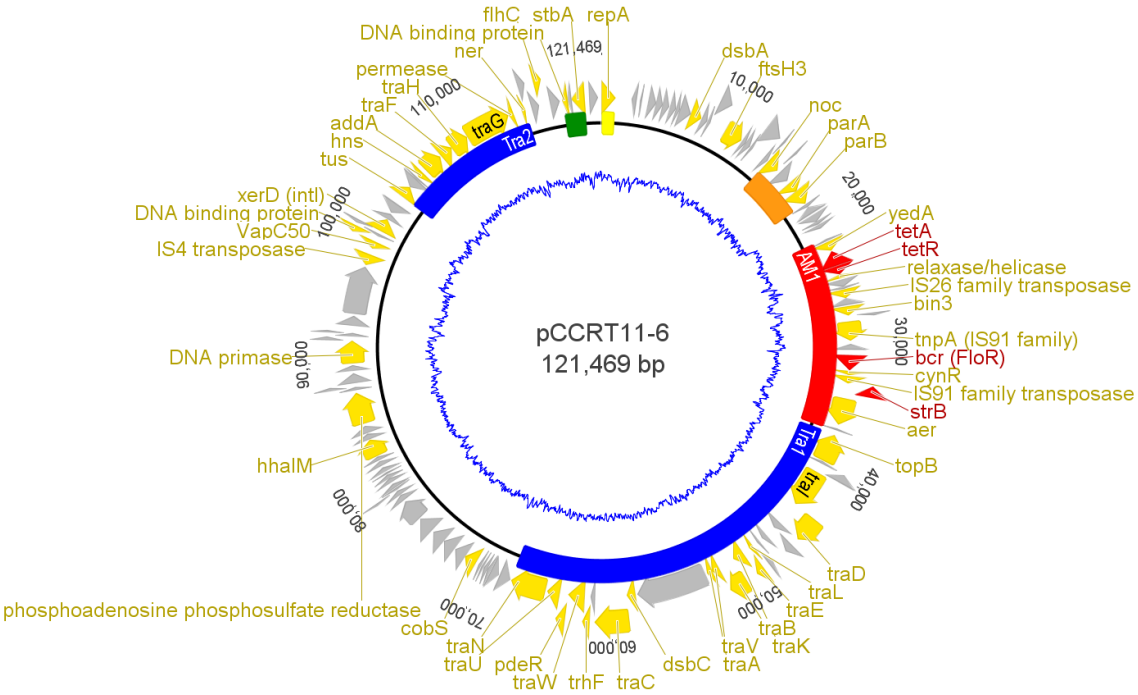

Figure 19. Functional modules and CDS mapped to the pCCRT11-6 genome assembly. Coding sequences and annotations are on the outer circle (outermost CDS are on complementary strand and the inner CDS are on the template strand): CDS annotated by Prokka and confirmed by BLASTp search (gold), antibiotic resistance genes annotated with ABRicate (red), and CDS annotated by Prokka as "hypothetical protein" (grey). Plasmid functional modules labeled on the inner circle: Accessory Module (AM1) (red), stability and partitioning regions (orange), conjugative transfer regions (Tra1/2) (blue), and replication regions (yellow). Inner blue trace represents % GC. pCCRT11-6 has a 48.89 % mean GC content.

The pCCRT11-6 genome includes 6 functional modules: replication, partitioning, accessory module 1 (AM1), 2 conjugative transfer (Tra) modules, and a stability module (Figure 19). pCCRT11-6 is likely a self-transmissible plasmid due to its ability to be conjugated into LA61 and the presence of the Tra modules that are typical in IncA/C_2 plasmids (Harmer and Hall, 2015). Both Tra modules contain genes that are involved in the type IV secretion system required for conjugation. The Tra1 module is a ~30 kb

region that consists of the genes *topB*, *traI/D/L/E/B/K/A/V*, *dsbC*, *traC*, *trhF*, *traW*, *pdeR*, and *traU/N*. The Tra2 module is a ~12 kb region that consists of the genes *tus*, *hns*, *addA*, *traF/H/G*, permease, and *ner*. IncA/C_2 conjugative transfer proteins have not been characterized experimentally; however the genes do share a high degree of similarity with related conjugative transfer genes from plasmids of other incompatibility groups such as certain IncF plasmids (Hammer and Hall, 2015). The Tra genes present on pCCRT11-6 align with a high degree of similarity (>82% nucleotide identity) to that of plasmid R16a and share nearly identical gene synteny (Figure 20a and 20b).

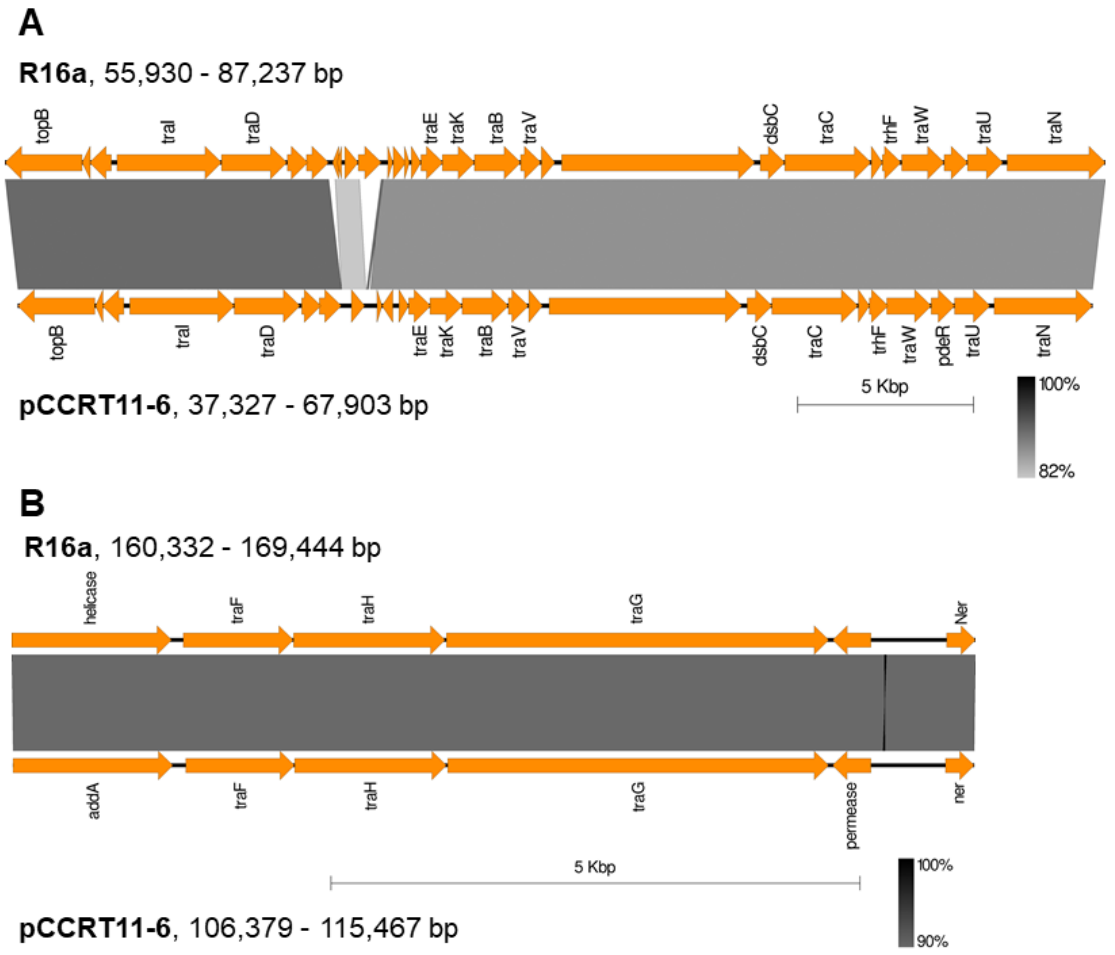

Figure 20. Comparison of the two conjugative transfer (Tra) modules of pCCRT11-6 with those of plasmid R16a. (A) The ~30 kb Tra1 region and (B) the ~12kb Tra2 region aligned to that of plasmid R16a. Orange arrows represent CDS and they point in the direction of transcription. CDS without labels are "hypothetical proteins." Alignment percent identity is shown by the key in the bottom right corner, with darker shades indicating a higher percent identity.

All IncA/C_2 plasmids have essential backbone genes involved in replication, maintenance, and stability (including plasmid partitioning), conjugative transfer, protein folding, and restriction/modification (Frick *et al.*, 2009). pCCRT11-6 contains all of the essential backbone genes including *repA* (replication initiation protein A), *tus* (DNA replication terminus site-binding protein), *parA/B* (plasmid partitioning proteins), *stbA* (essential for accurate partitioning of low-copy plasmids), conjugative transfer genes present in tra1 and tra2 modules, *hhaIM* (cytosine-specifc DNA methylase), *topB* (DNA

topoisomerase III), primase/helicase and *addA* (DNA primase/helicases), *dsbA/C* (disulphide bond proteins for ensuring correct protein folding) (Fricke *et al.*, 2009; Harmer and Hall, 2015).

In *E. coli*, *dsbA* is known to encode a protein that forms disulphide bonds as nascent peptides emerge into the cell's periplasm and dsbC encodes a protein that stabilizes disulphide bonds produced by *dsbA* (Harmer and Hall, 2015). Due to these genes location in the IncA/C_2 plasmid genomes, either upstream (*dsbA*) or within (*dsbC*) conjugative transfer modules, they are thought to stabilize and ensure the correct folding of the components of the Type IV secretion system used during conjugation (Harmer and Hall, 2015). The *dsbA/C* genes are both present within pCCRT11-6 and likely aid in successful conjugation and transfer of the plasmid into new hosts.

pCCRT11-6 resistance phenotype and genotype. Plasmid pCCRT11-6 conferred resistance to tetracycline (Table 7). This phenotype is likely due to the presence of the tetracycline resistance genes *tetA/R* that are located within accessory module 1 (AM1, Figure 19). Accessory module 1 also shows a % GC content that is higher and is a derivation from the mean 48.89% GC content, suggesting incorporation of this region by horizontal gene transfer in recent evolutionary history (Figure 19). Additional antibiotic and toxin resistance genes located within the genome, that were not tested phenotypically, were the streptomycin/spectinomycin resistance gene *strB*, the florfenicol resistance gene *bcr* (also known as *floR*), and the putative drug transporter gene *yedA* (also known as *eamA*) which is known to confer resistance to the toxic compound bromoacetate (Desai and Miller, 2010).

IncA/C plasmids are associated with the spread of genes conferring resistance to clinically relevant antibiotics such as carbapenems and third-generation cephalosporins (Harmer and Hall, 2015) however no β-lactamase resistance genes were found in pCCRT11-6. pCCRT11-6 does not have either anbiotic resistance island A (ARI-A) or ARI-B that are typically located in IncA/C plasmids. ARI-B is typically located directly upstream of the *parA/B* operon and contains antibiotic resistance genes, a class 1 integrons, and transposons however none were identified in that location in the pCCRT11-6 genome (Harmer and Hall, 2015). ARI-A is typicaly located directly upstream of the *rhs1* and *intI* genes and

Plasmids pCCP1 and pCCP2. The 58,942 bp pCCP1 and pCCP2 genomes are comprised of 71 CDS (34 labeled as "hypothetical protein") and 72 CDS (35 labeled as "hypothetical protein"), respectively (Tables S3, S4). The two plasmid genomes are nearly identical to one another, with evidence of a rearrangement of transposase-related CDS present in accessory module 1 (Figure 21).

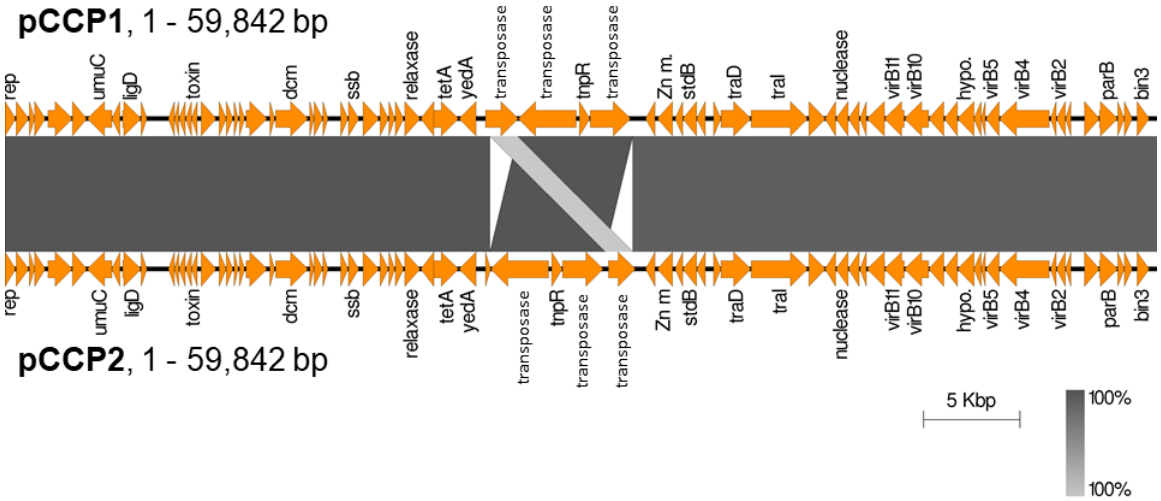

Figure 21. Comparison of pCCP1 and pCCP2 genomes. Orange arrows represent CDS and they point in the direction of transcription. CDS labeled as "hypo." or without labels are "hypothetical proteins." Alignment percent identity is shown by the key in the bottom right corner, with darker shades indicating a higher percent identity.

Both plasmids belong to the incompatibility group IncP-9 due to their similarity to the backbone regions (specifically the *rep* genes) of the IncP-9 plasmids pMT2 and pWW0 (Figure 22 and 24A, respectively). Sections of pCCP1 and pCCP2 align to pMT2, which is a ~10 kb plasmid that contains a cluster of replication and plasmid stability genes cloned into a new plasmid vector that were originally from the 75 kb IncP-9 plasmid pM3 (Sevastsyanovich *et al.*, 2005). The *repA* gene aligns with > 92% identity to that of pMT2, as well as the adjacent genes, *trbD/C*, *parA/B/R/C*, DNA resolvase, and an open reading frame (Figure 22).

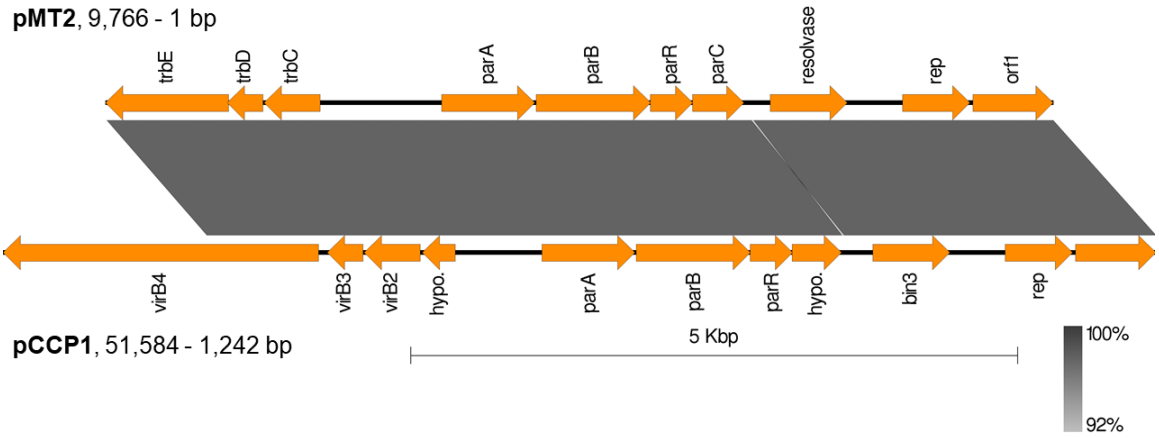

Figure 22. Comparison of the partitioning and replication regions of pCCP1 to the pMT2 plasmid. Orange arrows represent CDS and they point in the direction of transcription. CDS labeled as "hypo." or without labels are "hypothetical proteins." Alignment percent identity is shown by the key's in the bottom right corner, with darker shades indicating a higher percent identity.

The pCCP1 and pCCP2 genomes include 7 functional modules: replication, accessory module 1 (AM1), stability, DNA transfer (Dtr), mate pair formation (Mpf), and partitioning (Figure 23; pCCP2 map can be seen in supplemental materials Figure S20). Both pCCP1 and pCCP2 are likely self-transmissible plasmids due to their ability to transfer into LA61 and the presence of the Mpf and Dtr modules. The ~8.5 kb Mpf module consists of the genes *virB1/11/10/9/8*, type VI secretion protein-encoding gene, and *virB5/4/3/2* that encode proteins that are required for the synthesis of the type IV secretion system used during conjugation (Greated *et al.*, 2002). The ~10.3 kb Dtr module consists of the genes *traD*, *traD* (*trwB*), *traI*, *pld*, and putative nuclease that encode proteins that are necessary for conjugal DNA transfer (Greated *et al.*, 2002). Both the Mpf and Dtr modules align with > 84% identity to that of pWW0 which is an IncP-9 archetype ~117 kb plasmid with a ~46kb backbone originally isolated from *Pseudomonas putida* strain MT-2 (Figure 24b).

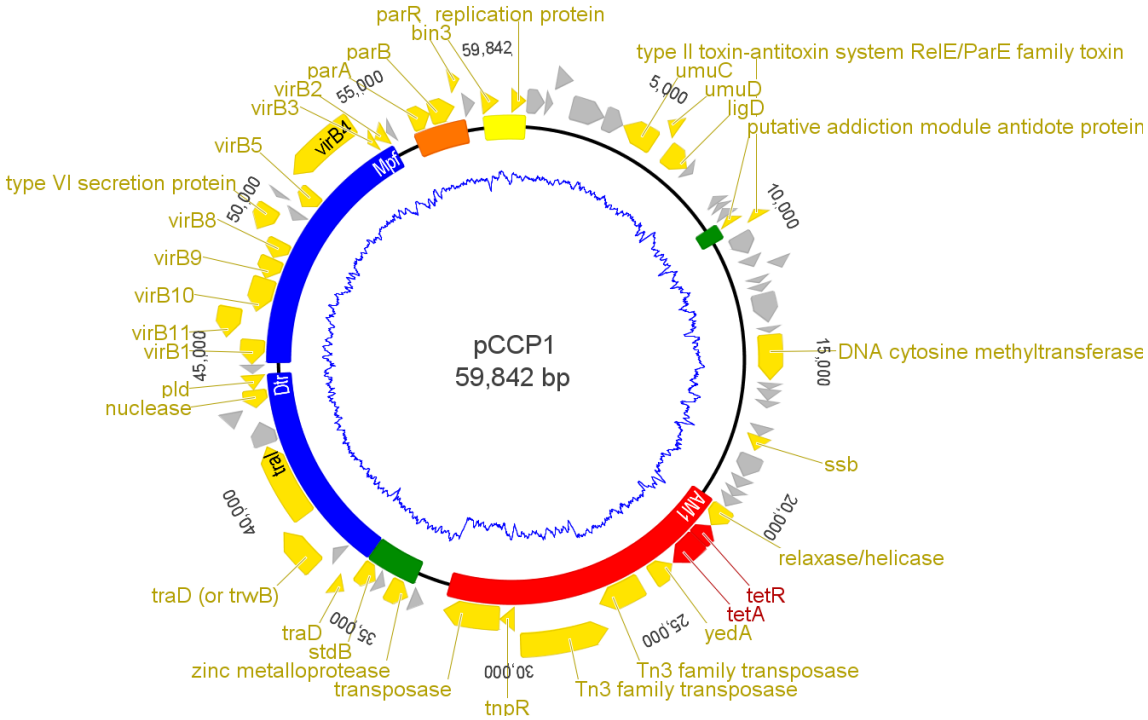

Figure 23. Functional modules and CDS mapped to the pCCP1 genome assembly. Coding sequences and annotations are on the outer circle (outermost CDS are on complementary strand and the inner CDS are on the template strand): CDS annotated by Prokka and confirmed by BLASTp search (gold), antibiotic resistance genes annotated with ABRicate (red), and CDS annotated by Prokka as "hypothetical protein" (grey). Plasmid functional modules labeled on the inner circle: Accessory Module (AM1) (red), stability and partitioning regions (orange), DNA transfer (Dtr) and mate pair formation (Mpf) regions (blue), and replication regions (yellow). Inner blue trace represents % GC. pCCP1 has a 57.04% mean GC content.

pCCP1 and pCCP2 stability and partitioning modules. The ~2.1 kb partitioning modules contain essential plasmid partitioning genes *parA/B/R* and an adjacent hypothetical protein that aligned with > 92% identity to the *parA/B/R/C* genes of pMT2 and with > 84% identity to the *parA/B* and *korA* genes of pWW0 (Figure 22 and 24b). The ~0.6 kb and ~2 kb stability modules contain a putative toxin-antitoxin system of the *RelE/ParE* family as well as *stdB* (plasmid stabilization protein) and a putative zinc metalloprotease.

pCCP1 and pCCP2 accessory modules. Both pCCP1 and pCCP2 contain ~12 kb accessory modules that contain the same genes, but with different synteny. These modules contain non-essential genes including a gene encoding for a putative DNA relaxase/helicase, tetracycline resistance genes *tetR/A*, putative drug transporter gene *yedA*, a hypothetical protein, Tn3 transposase, *tnpR*, and two additional transposase genes (Figure 23). The accessory modules also display lower % GC than the mean ~ 57% GC content of the entire genome, suggesting that the accessory modules was acquired by horizontal gene transfer in recent evolutionary history.

We hypothesize that these two slightly different plasmids may have originated from a single plasmid and diverged into two slightly different plasmids over the course of this study. The original host (an unknown bacterium from stream sediment at Cooks Creek Park) of this plasmid may have conjugated with multiple LA61 cells during the exogenous plasmid capture, giving rise to the growth of two distinct transconjugant colonies harboring identical tet^R plasmids. After extracting each of the plasmids, electrotransforming them into EC100, and maintaining the plasmids in separate EC100 cultures throughout this study, a transposon may have been horizontally transferred from the chromosome of EC100 into the backbone of the original plasmid, but in different ways between the two separate cultures. In pCCP1, the ~6 kb region from 26,600 – 32,581 bp aligns with 100% identity to a region of the EC100 genome containing the genes *tnpA* (transposase), *tnpR* (resolvase), and *tnpX* (encodes a protein of unknown function) which make up the Tn1000 transposon γδ (accession CP000948, E. coli str. K12 substr. DH10B, complete genome) (Wang *et al.*, 1994). In pCCP2, the ~6 kb region

from 25,165 – 31,157 bp aligns with 99.99% identity to the same region of the EC100 chromosome.

We hypothesize that the Tn1000 transposon γδ was horizontally transferred from EC100 into each of these identical tet^R plasmids at slightly different locations within the plasmids. Such artefacts caused by Tn1000 occur frequently in studies that use *E. coli* as host strain for cloning experiments due to its lack of sequence specificity for insertion (Broom *et al.*, 1995). This would explain why the two plasmids appear to be slightly different based on the restriction digest fragment patterns when each pCCP1 and pCCP2 were digested with EcoRI, SmaI, and HindIII (Supplementary Figure S19). Interestingly, Tn1000 has been used to study the closest related plasmid of pCCP1 and pCCP2, IncP-9 plasmid pWW0 for purposes of transposon mutagenesis (Harayama *et al.*, 1984). It is also possible that these assembly differences were caused by shared regions of sequence between the *E. coli* chromosome and the plasmids, however this would not explain the differences observed in the *in vitro* and *in silico* restriction digest patterns. Long-range PCR amplification of the accessory module regions of pCCP1 and pCCP2 followed by Sanger sequencing of the PCR products may help differentiate the differences between these nearly identical plasmids.

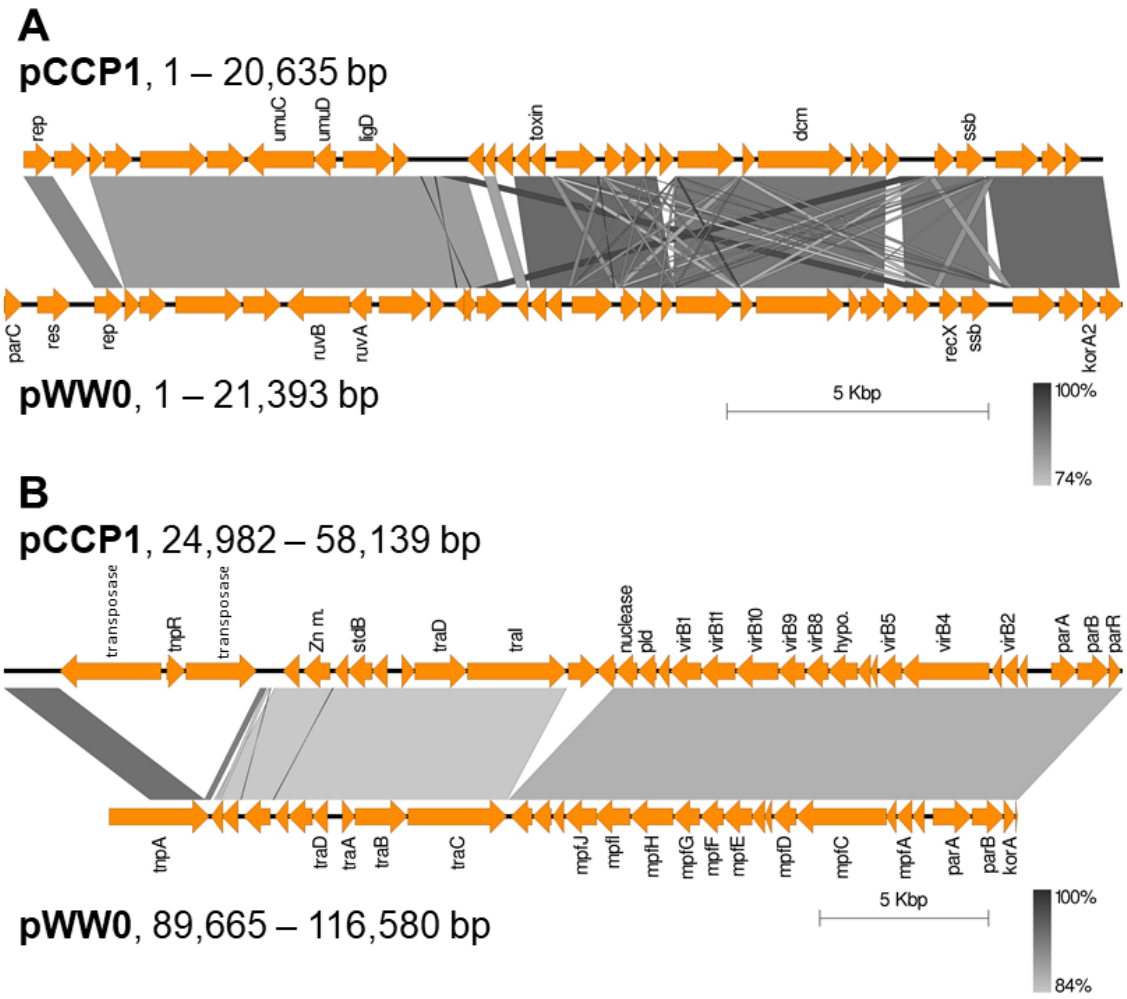

Figure 24. Comparison of the (A) beginning and (B) end of pCCP1 genome to IncP-9 plasmid pWW0. Orange arrows represent CDS and they point in the direction of transcription. CDS labeled as "hypo." or without labels are "hypothetical proteins." Alignment percent identity is shown by the key's in the bottom right corner, with darker shades indicating a higher percent identity

pCCP1 and pCCP2 resistance phenotypes and genotypes. Plasmids pCCP1 and pCCP2 conferred resistance to tetracycline (Table 7). This phenotype is likely due to the tetracycline resistance genes *tetA/R* that are located within accessory module 1 (AM1, Figures 22 and S20). No other antibiotic resistance genes were identified within the genomes, however a gene encoding a putative drug transporter *yedA*, which is known to confer resistance to bromoacetate, was identified. It was not tested phenotypically. Plasmis pCCP1 and pCCP2 also contain *umuC/D* (also known as *ruvB/A*) which confer

resistance to UV light, These genes are typical to find on IncP-9 plasmids (Greated *et al.*, 2002). UV resistance genes may be part of a repair process for damaged DNA.

Potential origins of pCCP1 and pCCP2. The pCCP1 and pCCP2 genomes have a mean % GC of 57.04 % and 57.09%, respectively. This is similar to the 57.8 % GC of the backbone regions of the IncP-9 plasmid pWW0 (Greated *et al.*, 2002). pWW0 was originally isolated from *Pseudomonas putida* strain mt-2, whose genome has a % GC of ~60%. This suggests that pCCP1 and pCCP2 may have originated from a *Pseudomonas* host. Plasmid pWW0 is a catabolic plasmid, with genes that code for the degradation and utilization of xylenes and toluenes, and most, if not all, IncP-9 plasmids have originated from *Pseudomonas* species and contain genes with catabolic functions (Dennis, 2005). Plasmids pCCP1 and pCCP2 do not contain any obvious catabolic genes, suggesting that any catabolic genes previously located on the genomes may have been lost in recent evolutionary history.

Conclusion

The pEG1-06, pCCRT11-6, pCCP1, and pCCP2 genomes demonstrate how genes conferring resistance to antibiotics may have been acquired by recent incorporation of integrons mobile genetic elements such as transposons or insertion sequences. Their genomes have a modular organization of essential backbone functions such as plasmid replication, maintenance and stability, partitioning, conjugative transfer as well as accessory modules containing antibiotic resistance genes and genes of unknown functions.

The ability of these four plasmids to be transferred into LA61 and the presence of the various conjugative transfer genes (Tra, Dtr, and Mpf regions) suggests that these plasmids are self-transmissible. The identification of the incompatibility groups of these plasmids, IncP-1β, IncA/C_2, and IncP-9 suggests that they are capable of horizontal transfer via conjugation and replication in a broad host range of bacteria. These plasmids have the potential to not only be transferred and replicated in a broad host range, but also have the potential to be stable and difficult to be lost or cured from the host due to the presence of the potential TA systems identified in each of the plasmids. Other factors also contribute to the stability of these plasmids such as the potential high copy number of pEG1-06.

Self-transmissible plasmids harboring multiple antibiotic resistance genes have been characterized from environments such as wetlands downstream of wastewater treatment plants (Botts *et al*., 2017), within waste water treatment plants (Schluter *et al.*, 2007), and livestock such as poultry (Rozwandowicz *et al.*, 2018). Both IncA/C and IncP plasmids are commonly found in clinical isolates and are associated with the worldwide

spread of multi-drug resistance conferred by genes encoding extended-spectrum beta-lactamases, carbapenemases, and aminoglycoside resistance (Rozwandowicz *et al*., 2018). IncP plasmids are known to be hosted by pathogenic *Salmonella enterica Typhimurium* and *E. coli* and these plasmids carry genes that confer resistance to last-resort antibiotics such as colistin (Lu *et al.*, 2016; Liu *et al.*, 2017). Antibiotic resistance among pathogenic bacteria is an urgent problem currently faced by public health officials, and the problem is exacerbated by self-transmissible plasmids spreading among nosocomial pathogens leading to nearly untreatable infections (San Milan, 2018).

Future studies are warranted to determine the minute differences between pCCP1 and pCCP2 genomes. Most of the fragments predicted by the *in silico* digests matched fragments observed in the *in vitro* digests, however they could be futher verified by visualization with a pulsed-field gel electrophoresis apparatus to view the larger fragments (>10 kb) and a normal agarose gel with a shorter run time to view the smaller fragments.

Appendix

Supplemental Materials

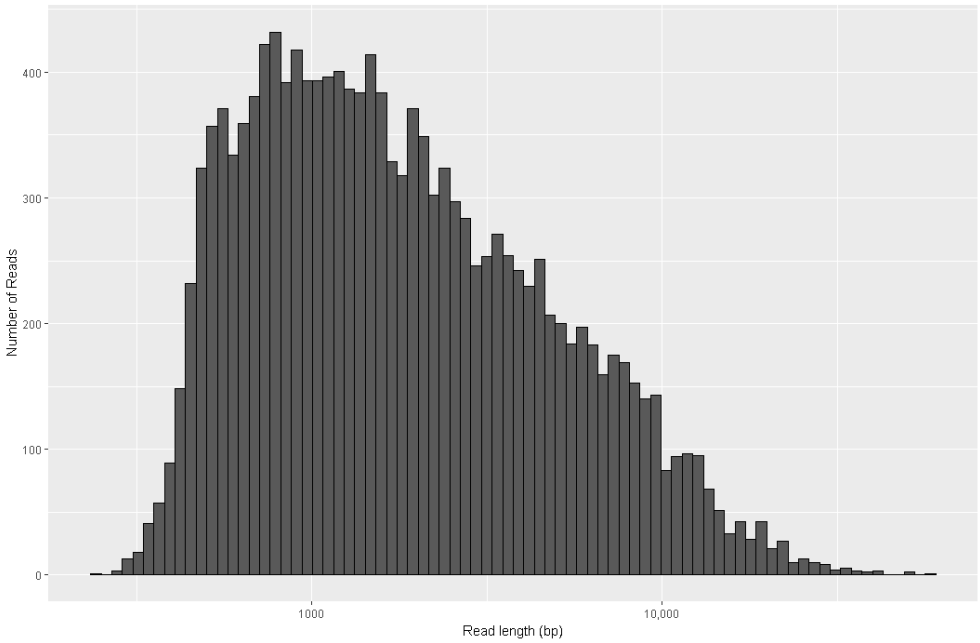

Figure S1. ONT read length distribution for pEG1-06 replicate 1. Mean read length was 2,958.7 bases and median read length was 1,550 bases. The longest read for this sample was 57,527 bases.

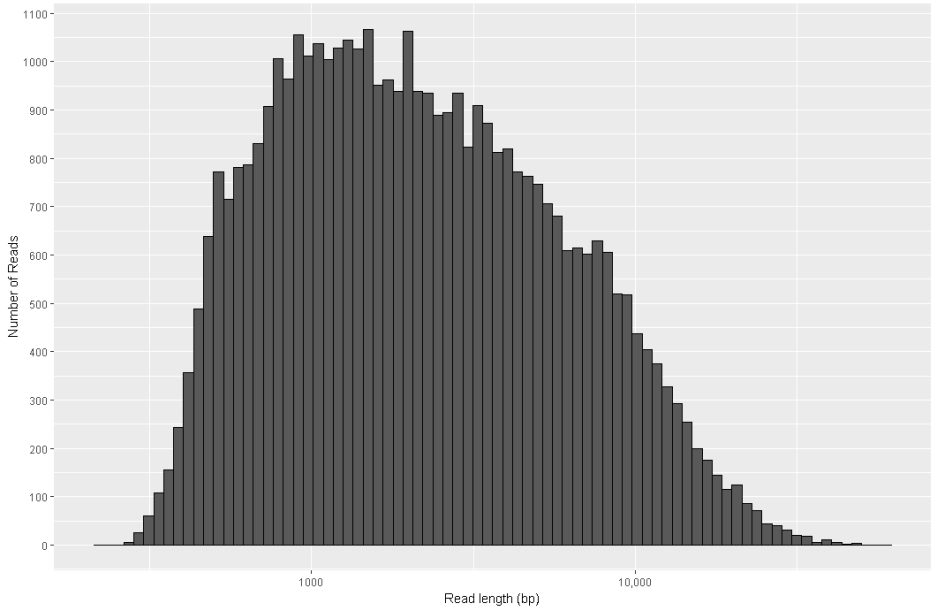

Figure S2. ONT read length distribution for pEG1-06 replicate 2. Mean read length was 3,275.7 bases and median read length was 1,972 bases. The longest read for this sample was 60,959 bases.

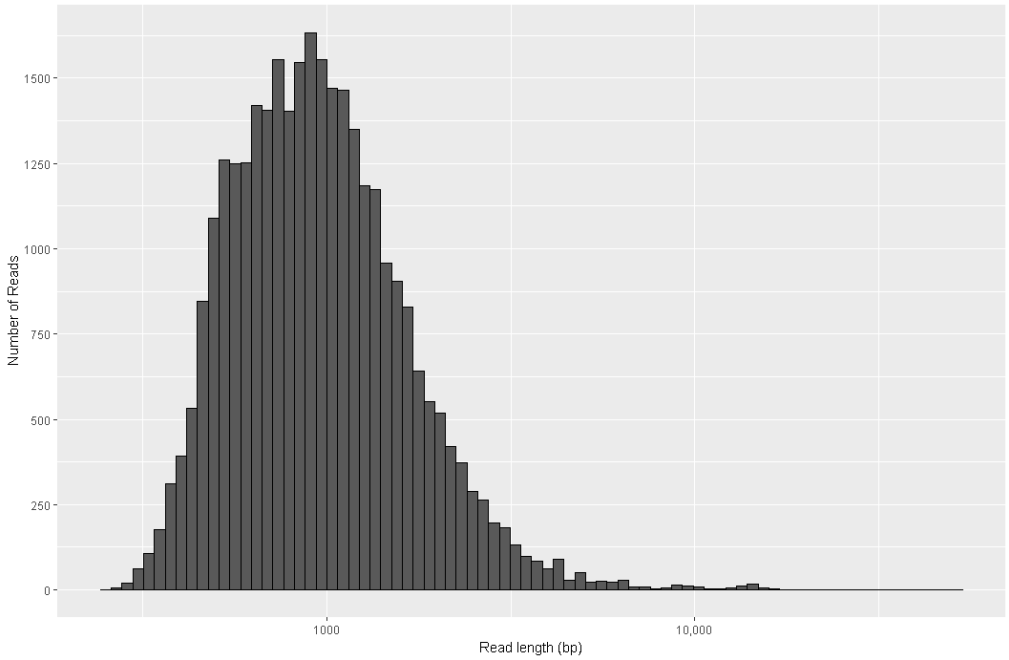

Figure S3. ONT read length distribution for pCCP1. Mean read length was 1,137 bases and median read length was 914 bases. The longest read for this sample was 52,991 bases.

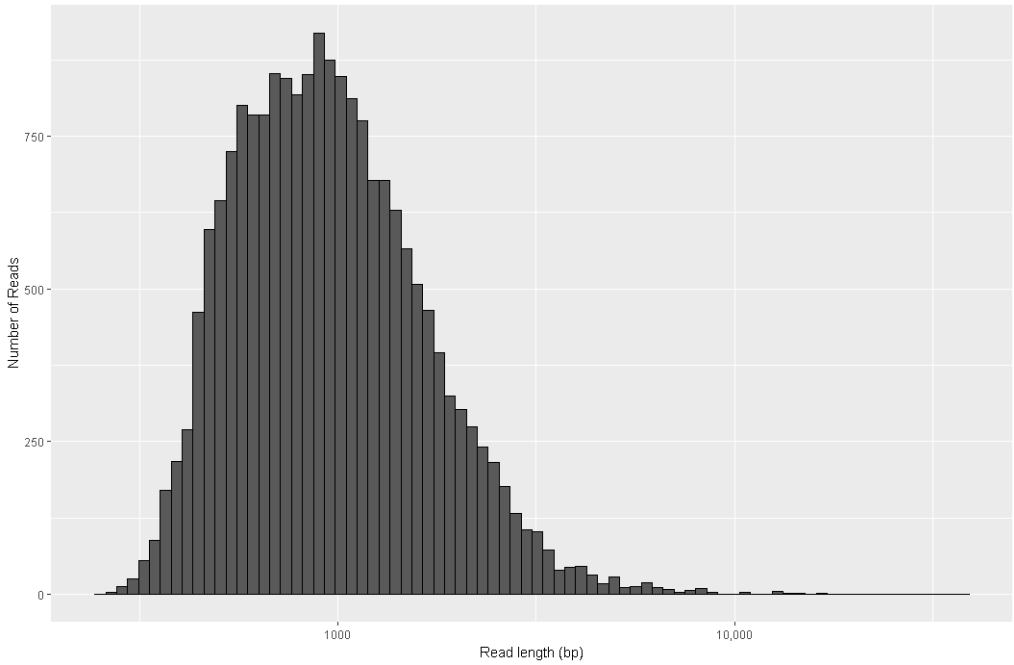

Figure S4. ONT read length distribution for pCCP2. Mean read length was 1,130.3 bases and median read length was 909 bases. The longest read for this sample was 37,146 bases.

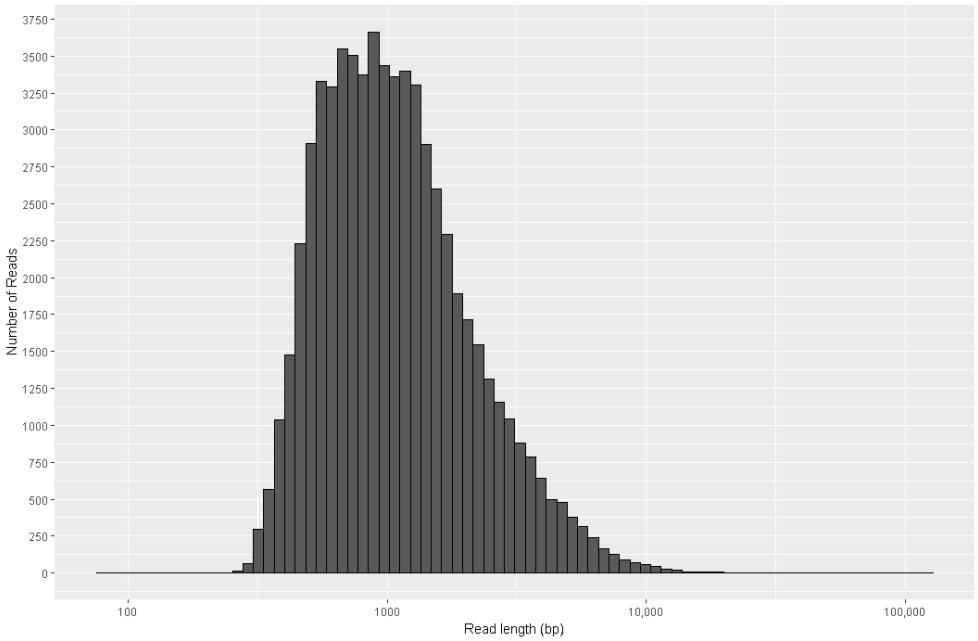

Figure S5. ONT read length distribution for pCCRT11-6. Mean read length was 1,399.1 bases and median read length was 998 bases. The longest read for this sample was 49,965 bases.

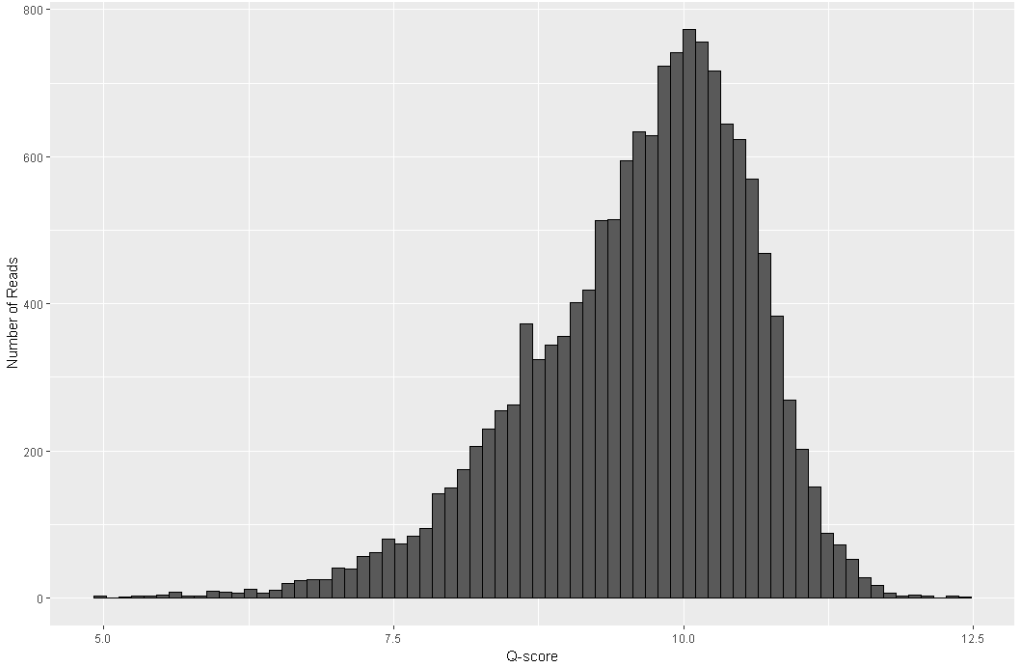

Figure S6. ONT read average Q score distribution for pEG1-06 replicate 1. Mean Q score was 9.6 and median Q score was 9.8. The highest Q score for this sample was 12.4.

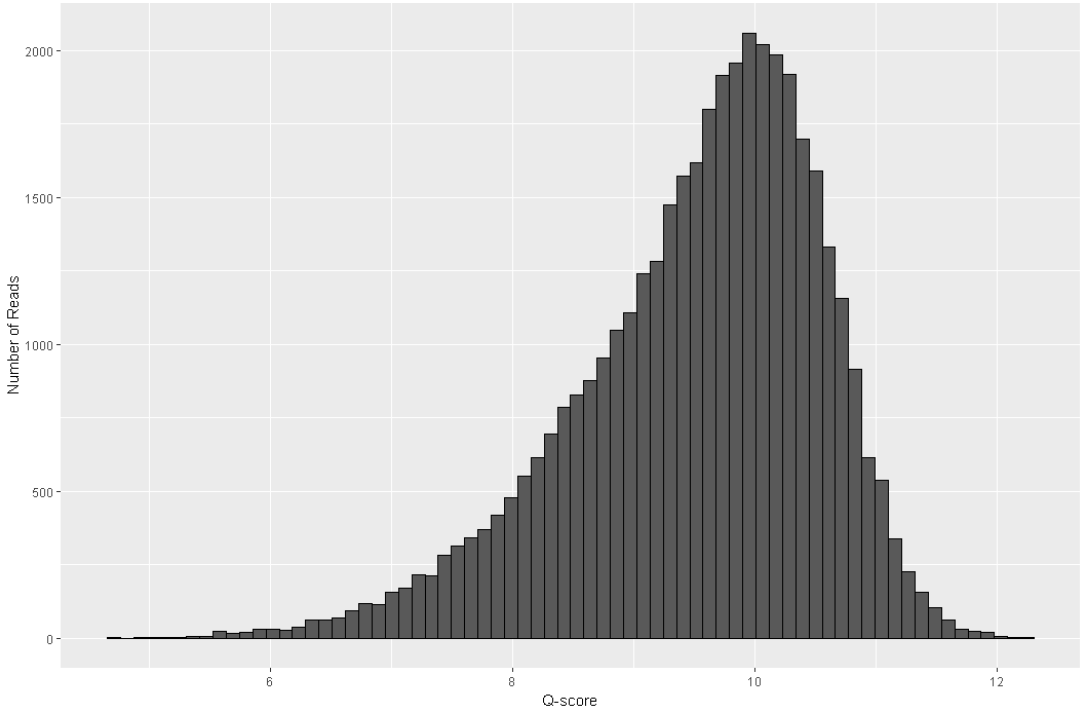

Figure S7. ONT read average Q score distribution for pEG1-06 replicate 2. Mean Q score was 9.5 and median Q score was 9.7. The highest Q score for this sample was 12.2.

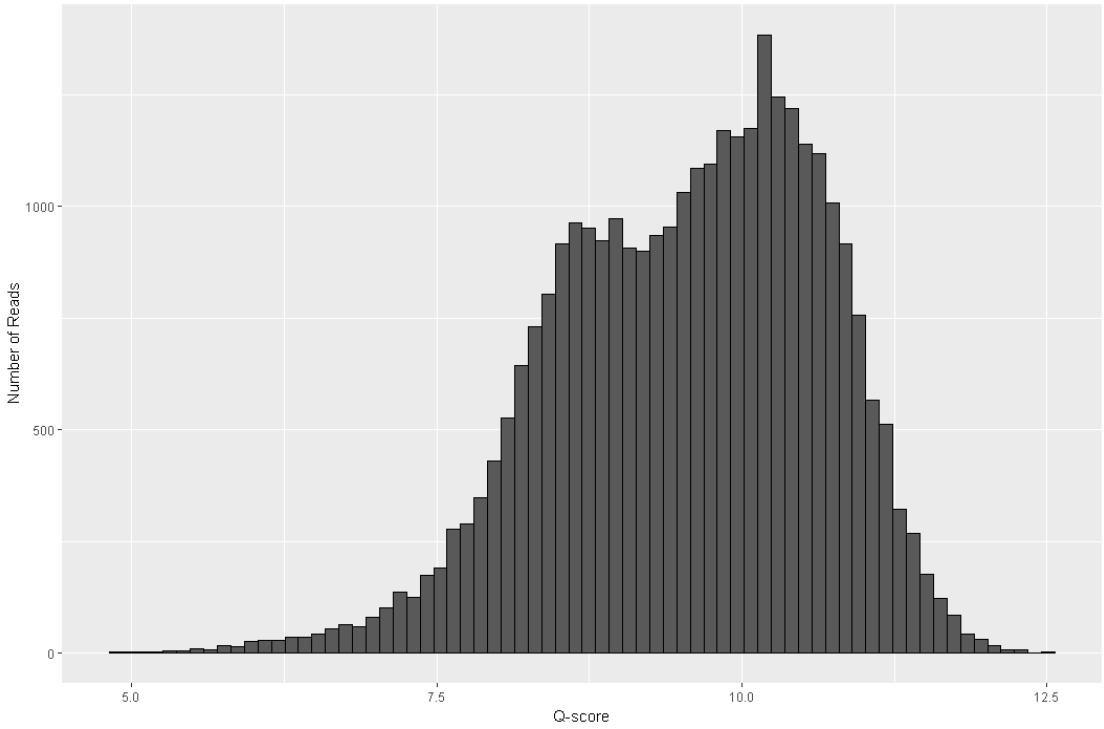

Figure S8. ONT read average Q score distribution for pCCP1. Mean Q score was 9.6 and median Q score was 9.8. The highest Q score for this sample was 12.5.

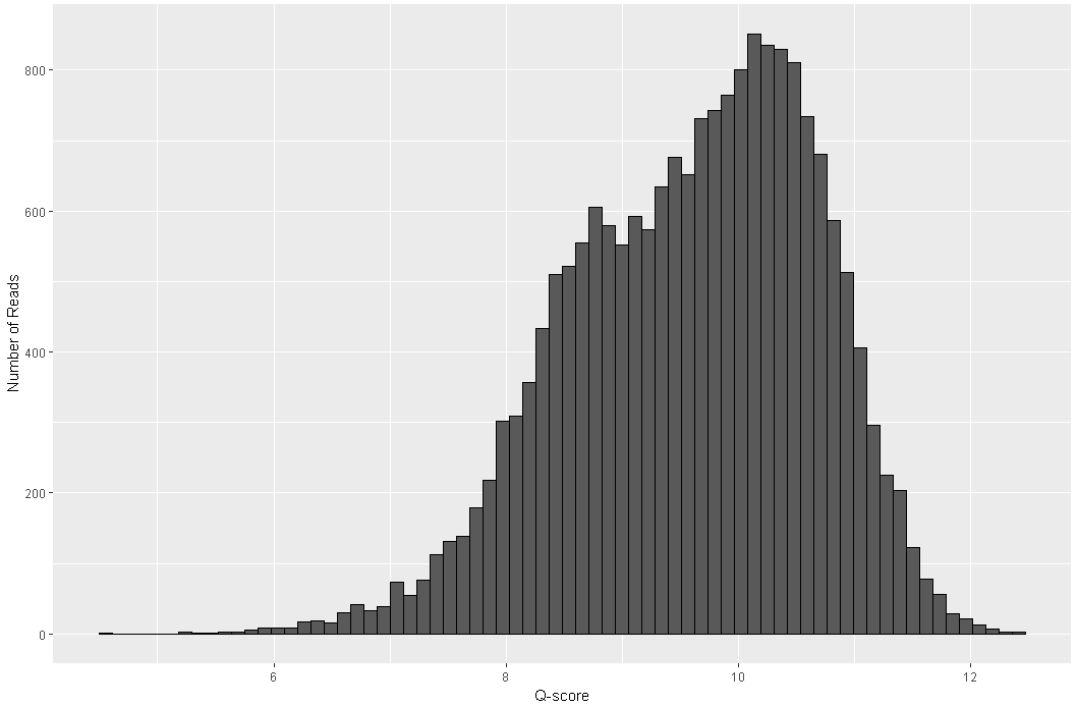

Figure S9. ONT read average Q score distribution for pCCP2. Mean Q score was 9.6 and median Q score was 9.7. The highest Q score for this sample was 12.4.

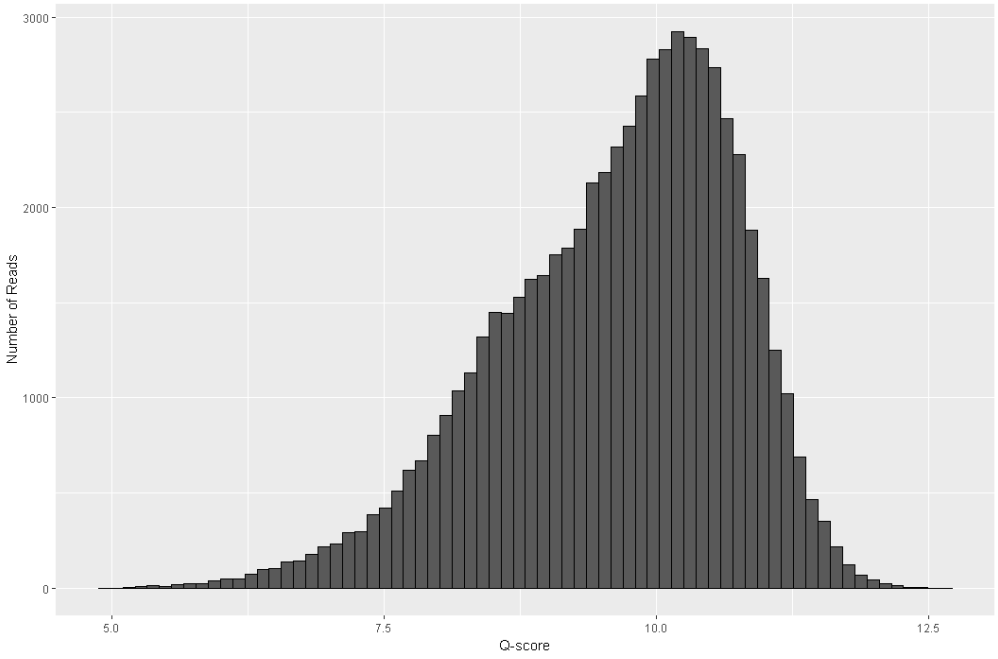

Figure S10. ONT read average Q score distribution for pCCRT11-6. Mean Q score was 9.6 and median Q score was 9.8. The highest Q score for this sample was 12.6.

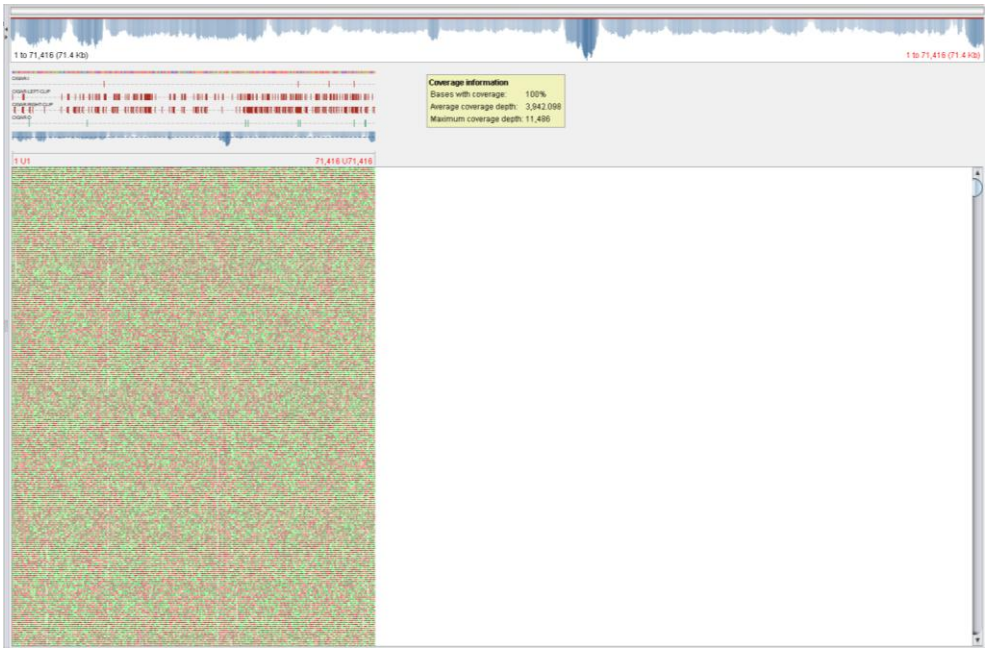

Figure S11. Tablet screenshot pEG1-06 Illumina reads aligned against the pEG1-06 assembly using BWA, visualized using Tablet (Li, 2018; Milne *et al.*, 2013).

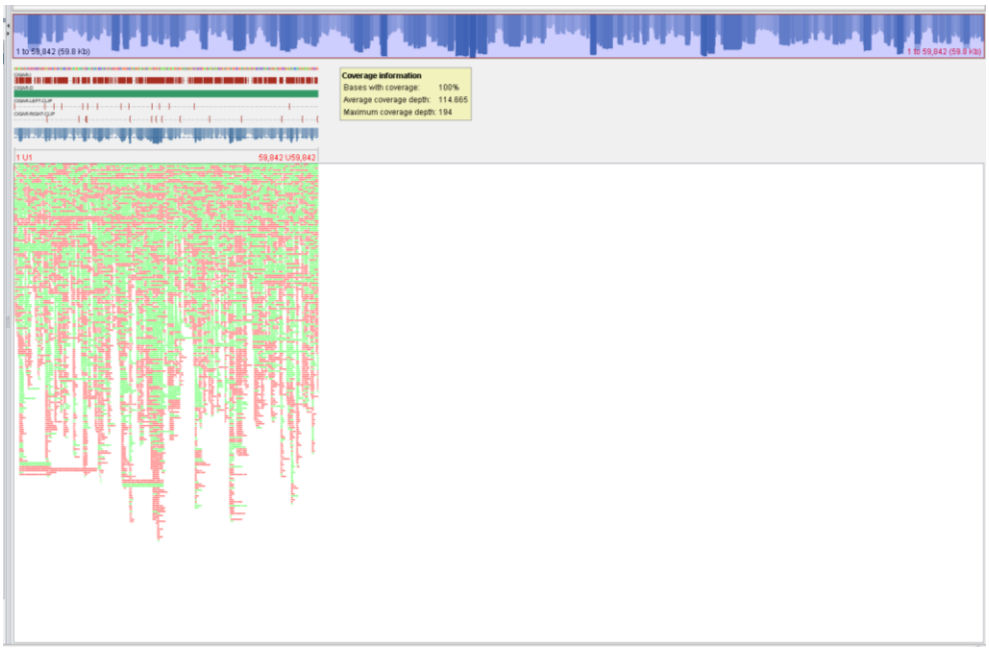

Figure S12. Tablet screenshot of ONT reads aligned to pCCP2 assembly using minimap2, visualized using Tablet (Li, 2018; Milne *et al.*, 2013).

Table S1. Complete list of pEG1-06 genes

Gene	Predicted function	Start	Stop	Length (bp)	Direction
strB	Aminoglycoside resistance	559	1,188	630	R
repA	DNA Replication	1,244	1,738	495	R
Hypothetical protein	Unknown	1,767	2,093	327	R
mazF	Toxin of toxin/antitoxin (TA) system	2,090	2,413	324	R
mazE	Antitoxin of TA system	2,413	2,637	225	R
DNA primase	Unknown	2,704	3,138	435	R
Hypothetical protein	Unknown	3,143	3,424	282	F
ofxX	Unknown	3,421	4,173	753	F
resolvase	Unknown	4,804	5,109	306	R
Hypothetical protein	Unknown	5,375	5,503	129	F
tetC	Tetracycline resistance	5,559	6,749	1,191	R
tetR	Tetracycline resistance	6,842	7,477	636	F
cytosine methyltransferase (IS200/IS605)	DNA methylation	7,500	7,835	336	R
transposase (IS605)	Transposase	8,098	9,411	1,314	F
Tn3 transposase	Transposase	9,514	10,734	1,221	F
trfA	Plasmid replication	11,357	12,577	1,221	R
ssb	Plasmid replication	12,624	12,965	342	R
trbA (trfB_1)	Mate pair formation (Mpf)	13,079	13,441	363	F
ptlH (trbB)	Mpf	13,751	14,713	963	F
trbC	Mpf	14,730	15,194	465	F
trbD	Mpf	15,198	15,509	312	F
virB4 (trbE)	Mpf	15,506	18,064	2,559	F
trbF	Mpf	18,061	18,843	783	F
virB9 (trbG)	Mpf	18,861	19,760	900	F
trbH	Mpf	19,763	20,251	489	F
virB10 (trbI)	Mpf	20,256	21,656	1,401	F
trbJ	Mpf	21,677	22,441	765	F
trbK	Mpf	22,451	22,678	228	F
trbL	Mpf	22,689	24,407	1,719	F
trbM	Mpf	24,425	25,012	588	F
trbN	Mpf	25,026	25,661	636	F
trbO	Mpf	25,690	25,956	267	F
trbP	Mpf	25,956	26,654	699	F

ompA	Mpf	26,670	27,101	432	F
DNA methylase	DNA methylation	27,256	27,930	675	F
xerD (intI)	Integrase	28,177	29,190	1,014	R
Hypothetical protein	Unknown	29,342	29,575	234	F
ant1_1 (ANT3)	Aminoglycoside resistance	29,648	30,487	840	F
emrE_1	Antiseptic resistance	30,651	30,998	348	F
folP_1 (sulI)	Sulfonamide resistance	30,992	31,963	972	F
ydhC (floR)	Florfenicol resistance	32,180	33,394	1,215	F
tetR	Tetracycline resistance	33,601	34,227	627	R
tetG	Tetracycline resistance	34,331	35,506	1,176	F
gltC	Transcriptional regulator	35,596	36,318	723	F
tnpA (IS91)	Transposase	36,410	37,942	1,533	R
groL	Molecular chaperone (protein folding)	38,169	38,822	654	R
pse1 (CARB-3)	Beta-lactam resistance	38,980	39,894	915	F
ant1_2 (ANT3)	Aminoglycoside resistance	39,994	40,803	810	F
emrE	Multi-drug efflux pump	40,967	41,314	348	F
folP_2 (sulI)	Sulfonamide resistance	41,308	42,147	840	F
N-acetyltransferase	Aminoglycoside resistance	42,275	42,775	501	F
tniB	Transposase	42,744	43,736	993	R
tnsB (tniA transposase)	Transposase	43,739	45,418	1,680	R
traC	DNA Transfer (Dtr)	45,969	50,315	4,347	R
traD	Dtr	50,319	50,708	390	R
topB (traE)	Dtr	50,730	52,793	2,064	R
traF	Dtr	52,805	53,341	537	R
traG	Dtr	53,338	55,251	1,914	R
traI	Dtr	55,248	57,488	2,241	R
traJ	Dtr	57,523	57,897	375	R
traK	Dtr	58,271	58,669	399	F
traL	Dtr	58,669	59,394	726	F
traM	Dtr	59,394	59,834	441	F
traN	Dtr	60,037	60,690	654	R
traB (traO)	Dtr	60,719	61,066	348	R
krfA	Plasmid maintenance/Stability	61,237	62,268	1,032	R
parB	Plasmid maintenance/Stability	62,448	63,497	1,050	R
parA	Plasmid maintenance/Stability	63,494	64,258	765	R

trfB (KorA)	Plasmid maintenance/Stability	64,255	64,566	312	R
kleF	Plasmid maintenance/Stability	64,671	65,201	531	R
kleE	Plasmid maintenance/Stability	65,203	65,532	330	R
kleB	Plasmid maintenance/Stability	65,677	65,892	216	R
kleA	Plasmid maintenance/Stability	65,951	66,187	237	R
korC	Plasmid maintenance/Stability, transcriptional regulator	66,347	66,604	258	R
klcB	Plasmid maintenance/Stability	66,621	67,826	1,206	R
klcA	Plasmid maintenance/Stability	67,878	68,306	429	R
nikR (type II TA system ParD family antitoxin)	Antitoxin of TA system	68,475	68,744	270	F
relE/ParE family toxin	Toxin of TA system	68,741	69,055	315	F
Hypothetical protein	Unknown	70,282	70,704	423	R
Hypothetical protein	Unknown	70,975	71,205	231	R
Hypothetical protein	Unknown	71,221	71,361	141	R

Table S2. Complete list of pCCRT11-6 genes

Gene	Predicted Function	Start	Stop	Length (bp)	Direction
repA	Plasmid replication	37	1,107	1,071	F
hypothetical protein	Unknown	2,333	2,767	435	R
hypothetical protein	Unknown	3,020	3,172	153	R
hypothetical protein	Unknown	3,674	4,282	609	F
hypothetical protein	Unknown	4,287	4,808	522	F
hypothetical protein	Unknown	4,811	5,332	522	F
hypothetical protein	Unknown	5,429	5,902	474	F
hypothetical protein	Unknown	5,913	6,206	294	F
hypothetical protein	Unknown	6,211	7,098	888	F
dsbA	Disulfide bond formation protein (protein folding stabilization)	7,114	7,971	858	F
hypothetical protein	Unknown	8,035	8,526	492	F
hypothetical protein	Unknown	8,538	8,765	228	F
hypothetical protein	Unknown	8,752	9,609	858	F
ftsH3	ATP-dependent zinc	10,071	11,534	1,464	F

	metalloprotease				
hypothetical protein	Unknown	11,609	11,938	330	F
hypothetical protein	Unknown	11,943	12,335	393	F
hypothetical protein	Unknown	12,384	12,581	198	F
hypothetical protein	Unknown	12,572	13,717	1,146	F
hypothetical protein	Unknown	13,729	14,076	348	F
hypothetical protein	Unknown	14,073	14,180	108	F
noc	Nucleoid occlusion protein	14,210	15,142	933	F
hypothetical protein	Unknown	15,142	15,321	180	F
hypothetical protein	Unknown	15,416	16,306	891	F
parA	Plasmid partitioning	16,570	17,355	786	F
parB	Plasmid partitioning	17,359	18,525	1,167	F
hypothetical protein	Unknown	18,574	18,846	273	F
hypothetical protein	Unknown	18,899	19,468	570	R
hypothetical protein	Unknown	19,613	20,188	576	F
hypothetical protein	Unknown	20,192	20,461	270	F
hypothetical protein	Unknown	20,495	20,776	282	F
hypothetical protein	Unknown	21,738	21,857	120	R
yedA	Multi-drug resistance (putative inner membrane transporter)	21,994	22,692	699	F
tetA	Tetracycline resistance	22,724	23,923	1,200	R
tetR	Tetracycline resistance	24,002	24,664	663	F
relaxase/helicase	Transposase	24,696	24,938	243	R
hypothetical protein	Unknown	25,086	25,661	576	F
hypothetical protein	Unknown	25,665	25,841	177	F
IS26 family transposase	Transposase	25,853	26,557	705	F
hypothetical protein	Unknown	26,591	27,082	492	R
bin3	Putative transposon Tn552 DNA-invertase (resolvase)	27,189	27,926	738	F
hypothetical protein	Unknown	27,923	28,147	225	F
tnpA (IS91 family transposase)	transposase	28,358	29,851	1,494	F
hypothetical protein	Unknown	30,149	30,604	456	R
bcr (FloR)	Bicyclomycin/chloramphenicol/florfenicol resistance	30,983	32,197	1,215	F
cynR	HTH-type transcriptional regulator	32,225	32,530	306	F

IS91 family transposase	transposase	32,642	33,181	540	F
strB	Aminoglycoside resistance	33,153	33,989	837	R
Aer	Aerotaxis receptor	34,364	36,472	2,109	F
hypothetical protein	Unknown	36,796	36,939	144	R
topB	DNA topoisomerase 3, conjugative transfer	37,327	39,516	2,190	R
hypothetical protein	Unknown	39,540	39,749	210	R
hypothetical protein	Unknown	39,731	40,348	618	R
traI	conjugative transfer (Tra)	40,516	43,491	2,976	F
traD	Tra	43,488	45,353	1,866	F
hypothetical protein	Unknown	45,403	45,948	546	F
hypothetical protein	Unknown	45,905	46,534	630	F
hypothetical protein	Unknown	46,819	47,196	378	F
hypothetical protein	Unknown	47,547	47,690	144	F
hypothetical protein	Unknown	47,687	47,998	312	R
traL	Tra	48,169	48,450	282	F
traE	Tra	48,450	49,073	624	F
traK	Tra	49,057	49,974	918	F
traB	Tra	49,971	51,287	1,317	F
traV	Tra	51,284	51,856	573	F
traA	Tra	51,869	52,252	384	F
hypothetical protein	Unknown	52,442	57,928	5,487	F
dsbC	Protein folding	58,076	58,783	708	F
traC	Tra	58,780	61,227	2,448	F
hypothetical protein	Unknown	61,242	61,559	318	F
trhF	Conjugative transfer signal peptidase	61,556	62,086	531	F
traW	Dtr	62,049	63,314	1,266	F
pdeR	Cyclic di-GMP phosphodiesterase	63,311	63,982	672	F
traU	Dtr	63,979	64,986	1,008	F
traN	Dtr	65,102	67,903	2,802	F
hypothetical protein	Unknown	67,991	68,851	861	R
hypothetical protein	Unknown	68,973	69,623	651	R
hypothetical protein	Unknown	69,680	69,910	231	F
hypothetical protein	Unknown	69,921	70,241	321	F
hypothetical protein	Unknown	70,336	70,428	93	F
hypothetical protein	Unknown	70,477	70,707	231	F
cobS	Aerobic cobaltochelatase subunit	70,947	71,915	969	F

hypothetical protein	Unknown	71,926	72,831	906	F
hypothetical protein	Unknown	72,956	73,921	966	F
hypothetical protein	Unknown	73,982	74,992	1,011	F
hypothetical protein	Unknown	75,191	76,456	1,266	F
hypothetical protein	Unknown	76,541	78,253	1,713	F
hypothetical protein	Unknown	78,368	78,856	489	F
hypothetical protein	Unknown	78,922	79,254	333	F
hypothetical protein	Unknown	79,247	79,369	123	R
hypothetical protein	Unknown	79,366	79,503	138	R
hypothetical protein	Unknown	79,574	79,927	354	F
hypothetical protein	Unknown	80,063	80,431	369	F
hypothetical protein	Unknown	80,443	80,718	276	F
hypothetical protein	Unknown	80,776	80,967	192	F
hypothetical protein	Unknown	81,036	81,389	354	F
hypothetical protein	Unknown	81,583	81,927	345	F
hypothetical protein	Unknown	82,004	82,306	303	F
hhaIM	Modification methylase HhaI	82,384	83,772	1,389	F
hypothetical protein	Unknown	83,833	84,264	432	F
hypothetical protein	Unknown	84,337	84,891	555	F
phosphoadenosine phosphosulfate reductase	Sulfur metabolism	85,121	87,571	2,451	F
hypothetical protein	Unknown	87,773	88,021	249	F
hypothetical protein	Unknown	88,018	88,230	213	F
hypothetical protein	Unknown	88,312	89,109	798	F
hypothetical protein	Unknown	89,312	89,698	387	F
DNA primase	Unknown	89,876	91,621	1,746	F
hypothetical protein	Unknown	91,608	91,868	261	F
hypothetical protein	Unknown	91,960	92,208	249	F
hypothetical protein	Unknown	92,205	92,324	120	F
hypothetical protein	Unknown	92,798	93,523	726	F
hypothetical protein	Unknown	93,703	97,461	3,759	F
IS4 transposase	transposase	97,604	98,491	888	R
VapC50	Putative ribonuclease (toxin) of TA system	98,961	99,536	576	R
DNA binding protein	Unknown	99,536	99,997	462	R
hypothetical protein	Unknown	99,994	100,233	240	R
hypothetical protein	Unknown	100,233	100,445	213	R
xerD	Tyrosine recombinase (integrase)	100,474	101,472	999	F
hypothetical protein	Unknown	101,529	102,023	495	F

hypothetical protein	Unknown	102,116	102,910	795	R
tus	DNA replication terminus site-binding protein	103,667	104,542	876	F
hypothetical protein	Unknown	104,555	105,340	786	R
hypothetical protein	Unknown	105,341	105,487	147	R
hns	DNA-binding protein	105,599	106,024	426	R
addA	ATP-dependent helicase/nuclease subunit A	106,379	107,890	1,512	F
traF	Tra	108,013	109,041	1,029	F
traH	Tra	109,043	110,476	1,434	F
traG	Tra	110,489	114,091	3,603	F
permease	Unknown	114,128	114,490	363	R
ner	Nucleotide excision repair protein	115,195	115,467	273	F
transglycosylase	Soluble lytic murein transglycosylase and related regulatory protein	115,467	116,000	534	F
hypothetical protein	Unknown	116,009	116,617	609	F
flhC	Flagellar transcriptional regulator	116,614	117,165	552	F
hypothetical protein	Unknown	117,472	118,245	774	F
DNA binding protein	Unknown	118,452	118,871	420	R
hypothetical protein	Unknown	118,873	119,166	294	R
stbA	Plasmid stability	119,183	120,166	984	R
hypothetical protein	Unknown	120,567	121,193	627	F

Table S3. Complete list of pCCP1 genes

gene	Predicted Function	Start	Stop	Length (bp)	Direction
replication protein	DNA replication	1	555	555	F
hypothetical protein	Unknown	580	1,242	663	F
hypothetical protein	Unknown	1,260	1,550	291	F
hypothetical protein	Unknown	1,543	2,088	546	F
hypothetical protein	Unknown	2,229	3,500	1,272	F
hypothetical protein	Unknown	3,502	4,242	741	F
umuC	Y-family DNA polymerase	4,263	5,549	1,287	R
umuD	UV resistance protein	5,539	5,970	432	R

ligD	Multifunctional non-homologous end joining protein	6,109	7,062	954	F
hypothetical protein	Unknown	7,073	7,366	294	F
hypothetical protein	Unknown	8,476	8,793	318	R
hypothetical protein	Unknown	8,795	9,016	222	R
hypothetical protein	Unknown	9,029	9,352	324	R
putative addiction module antidote protein	Antitoxin of TA system	9,375	9,674	300	R
type II toxin-antitoxin system RelE/ParE family toxin	Toxin of TA system	9,671	9,973	303	R
hypothetical protein	Unknown	10,177	10,956	780	F
hypothetical protein	Unknown	11,110	11,478	369	F
hypothetical protein	Unknown	11,478	11,828	351	F
hypothetical protein	Unknown	11,885	12,103	219	F
hypothetical protein	Unknown	12,164	12,445	282	F
hypothetical protein	Unknown	12,512	13,603	1,092	F
hypothetical protein	Unknown	13,750	13,992	243	F
DNA cytosine methyltransferase	DNA methylation	14,044	15,726	1,683	F
hypothetical protein	Unknown	15,820	16,041	222	F
hypothetical protein	Unknown	16,045	16,485	441	F
hypothetical protein	Unknown	16,493	16,759	267	F
hypothetical protein	Unknown	17,418	17,807	390	F
ssb	Single-stranded DNA-binding protein	17,832	18,365	534	F
hypothetical protein	Unknown	18,584	19,411	828	F
hypothetical protein	Unknown	19,470	19,901	432	F
hypothetical protein	Unknown	19,918	20,232	315	F
hypothetical protein	Unknown	20,252	20,686	435	F
relaxase/helicase	Unknown	20,742	21,545	804	F
tetR	Tetracycline repressor protein class A from transposon 1721	21,577	22,254	678	R
tetA	Tetracycline resistance protein, class C	22,258	23,532	1,275	F
yedA	putative inner membrane transporter	23,564	24,448	885	R
Tn3 family transposase	Transposase	24,937	26,637	1,701	F
Tn3 family	Transposase	26,634	29,642	3,009	R

transposase					
tnpR	Transposon gamma-delta resolvase	29,806	30,357	552	F
transposase	Transposase	30,373	32,469	2,097	F
hypothetical protein	Unknown	33,248	33,730	483	R
zinc metalloprotease	Unknown	33,849	34,646	798	R
hypothetical protein	Unknown	34,777	35,169	393	R
stdB	hypothetical protein	35,173	35,892	720	R
traD	Dtr	35,889	36,344	456	R
hypothetical protein	Unknown	36,776	37,159	384	F
traD (or trwB)	Dtr	37,159	38,709	1,551	F
traI	Dtr	38,721	41,657	2,937	F
hypothetical protein	Unknown	41,710	42,579	870	F
hypothetical protein	Unknown	42,551	43,108	558	R
nuclease	Unknown	43,120	43,755	636	R
pld	Phospholipase D (endonuclease)	43,768	44,319	552	R
hypothetical protein	Unknown	44,363	44,698	336	R
virB1	Type IV secretion system protein	44,727	45,644	918	R
virB11	Type IV secretion system protein	45,628	46,656	1,029	R
virB10	Type IV secretion system protein	46,666	47,928	1,263	R
virB9	Type IV secretion system protein	47,931	48,716	786	R
virB8	Type IV secretion system protein	48,744	49,418	675	R
type VI secretion protein	Type VI secretion protein	49,418	50,281	864	R
hypothetical protein	Unknown	50,269	50,661	393	R
hypothetical protein	Unknown	50,654	50,866	213	R
virB5	Type IV secretion system protein	50,895	51,587	693	R
virB4	Type IV secretion system protein	51,584	54,184	2,601	R
virB3	Type IV secretion system protein	54,255	54,551	297	R
virB2	Type IV secretion system protein	54,560	55,021	462	R
Hypothetical protein	Unknown	55,038	55,310	273	R
parA	Plasmids partitioning	56,024	56,791	768	F
parB	Plasmid partitioning	56,802	57,743	942	F
parR	Plasmid partitioning	57,740	58,090	351	F
hypothetical protein	Unknown	58,087	58,494	408	F

bin3 (resolvase)	Putative transposon Tn552 DNA-invertase (resolvase)	58,751	59,386	636	F

Table S4. Complete list of pCCP2 genes

Gene	Product	Start	Stop	Length (bp)	Direction
replication protein	Plasmid replication	1	555	555	F
hypothetical protein	Unknown	580	1,242	663	F
hypothetical protein	Unknown	1,260	1,550	291	F
hypothetical protein	Unknown	1,543	2,088	546	F
hypothetical protein	Unknown	2,229	3,500	1,272	F
hypothetical protein	Unknown	3,502	4,242	741	F
umuC	Protein UmuC	4,263	5,549	1,287	R
umuD	Protein UmuD	5,539	5,970	432	R
ligD	Multifunctional non-homologous end joining protein LigD	6,109	7,062	954	F
hypothetical protein	Unknown	7,073	7,366	294	F
hypothetical protein	Unknown	8,476	8,793	318	R
hypothetical protein	Unknown	8,795	9,016	222	R
hypothetical protein	Unknown	9,029	9,352	324	R
putative addiction module antidote protein	Antitoxin of TA system	9,375	9,674	300	R
type II toxin-antitoxin system RelE/ParE family toxin	Toxin of TA system	9,671	9,973	303	R
hypothetical protein	Unknown	10,177	10,956	780	F
hypothetical protein	Unknown	11,110	11,478	369	F
hypothetical protein	Unknown	11,478	11,828	351	F
hypothetical protein	Unknown	11,885	12,103	219	F
hypothetical protein	Unknown	12,164	12,445	282	F
hypothetical protein	Unknown	12,512	13,603	1,092	F
hypothetical protein	Unknown	13,750	13,992	243	F
DNA cytosine methyltransferase	Methylation	14,044	15,726	1,683	F
hypothetical protein	Unknown	15,820	16,041	222	F
hypothetical protein	Unknown	16,045	16,485	441	F
hypothetical protein	Unknown	16,493	16,759	267	F
hypothetical protein	Unknown	17,418	17,807	390	F

ssb	Single-stranded DNA-binding protein	17,832	18,365	534	F
hypothetical protein	Unknown	18,584	19,411	828	F
hypothetical protein	Unknown	19,470	19,901	432	F
hypothetical protein	Unknown	19,918	20,232	315	F
hypothetical protein	Unknown	20,252	20,686	435	F
relaxase/helicase	Unknown	20,742	21,545	804	F
tetR	Tetracycline repressor protein class A from transposon 1721	21,577	22,254	678	R
tetA	Tetracycline resistance protein, class C	22,258	23,532	1,275	F
yedA	putative inner membrane transporter YedA	23,564	24,448	885	R
hypothetical protein	Unknown	24,937	25,203	267	F
Tn3 family transposase	Transposase	25,200	28,208	3,009	R
tnpR	Transposon gamma-delta resolvase	28,372	28,923	552	F
transposase	Transposase	28,939	31,035	2,097	F
Tn3-like element TnAs1 family transposase	Transposase	31,310	32,719	1,410	F
hypothetical protein	Unknown	33,248	33,730	483	R
zinc metalloprotease	Unknown	33,849	34,646	798	R
hypothetical protein	Unknown	34,777	35,169	393	R
stdB	Plasmid stability	35,173	35,892	720	R
traD	DNA Transfer (Dtr)	35,889	36,344	456	R
hypothetical protein	Unknown	36,776	37,159	384	F
traD (or trwB)	Dtr	37,159	38,709	1,551	F
traI	Dtr	38,721	41,657	2,937	F
hypothetical protein	Unknown	41,710	42,579	870	F
hypothetical protein	Unknown	42,551	43,108	558	R
nuclease	Unknown	43,120	43,755	636	R
pld	Phospholipase D	43,768	44,319	552	R
hypothetical protein	Unknown	44,363	44,698	336	R
virB1	Type IV secretion system protein	44,727	45,644	918	R
virB11	Type IV secretion system protein	45,628	46,656	1,029	R
virB10	Type IV secretion	46,666	47,928	1,263	R

	system protein				
virB9	Type IV secretion system protein	47,931	48,716	786	R
virB8	Type IV secretion system protein	48,744	49,418	675	R
type VI secretion protein	Type VI secretion system protein	49,418	50,281	864	R
hypothetical protein	Unknown	50,269	50,661	393	R
hypothetical protein	Unknown	50,654	50,866	213	R
virB5	Type IV secretion system protein	50,895	51,587	693	R
virB4	Type IV secretion system protein	51,584	54,184	2,601	R
virB3	Type IV secretion system protein	54,255	54,551	297	R
virB2	Type IV secretion system protein	54,560	55,021	462	R
hypothetical protein	Unknown	55,038	55,310	273	R
parA	Plasmid partitioning	56,024	56,791	768	F
parB	Plasmid partitioning	56,802	57,743	942	F
parR	Plasmid partitioning	57,740	58,090	351	F
hypothetical protein	Unknown	58,087	58,494	408	F
bin3	Putative transposon Tn552 DNA-invertase (resolvase)	58,751	59,386	636	F

Table S5. *In silico* digest predictions and observations of pEG1-06 digested with EcoRI

Fragment number	Fragment Size (bp)	Seen on gel image?
F4	25,817	Yes, but bands overlap
F2	21,623	Yes, but bands overlap
F1	12,010	Yes
F6	8,331	Yes
F3	2,772	Yes
F5	863	Yes

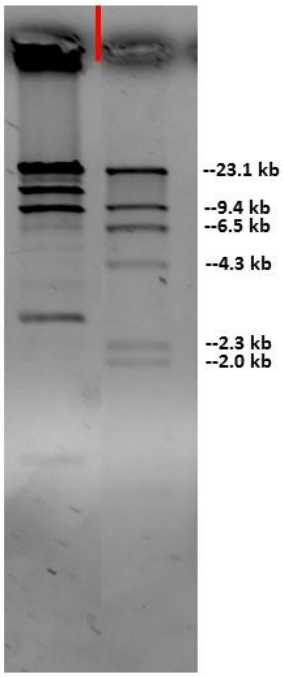

Figure S13. Restriction digestion of pEG1-06 with EcoRI. Left lane: pEG1-06 digested by EcoRI, right lane: lambda/HindIII molecular weight standard. Red line indicates where the image was cut to remove unnecessary lanes.

Table S6. *In silico* digest predictions and observations of pEG1-06 digested with HindIII

Fragment number	Fragment Size (bp)	Seen on gel image?
F8	28,830	Yes, but bands overlap
F2	19,889	Yes, but bands overlap
F4	9,286	Yes
F9	8,177	Yes
F1	4,045	Yes
F5	717	No
F7	159	No
F3	159	No
F6	154	No

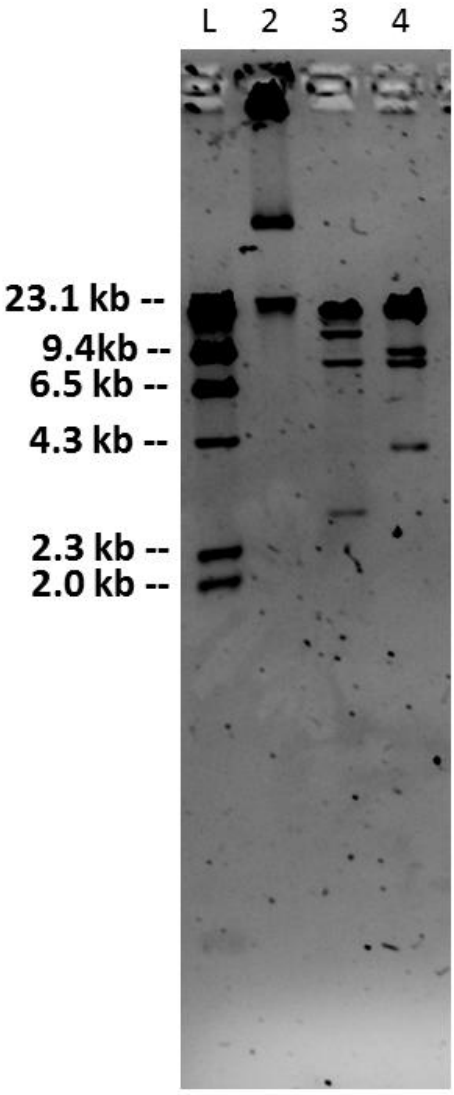

Figure S14. Restriction digestion of pEG1-06 with EcoRI and HindIII. Lane L: lambda/HindIII digest, lane 2: pEG1-06, lane 3: pEG1-06 digested with EcoRI, lane 4: pEG1-06 digested with HindIII.

Table S7. *In silico* digest predictions and observations of pEG1-06 digested with SmaI

Fragment number	Fragment Size (bp)	Seen on gel image?
F3	43,580	Yes
F2	17,130	Yes
F1	10,706	Yes

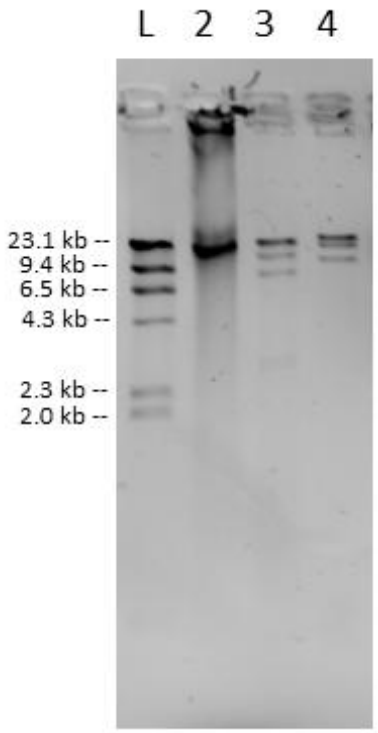

Figure S15. Restriction digestion of pEG1-06 with EcoRI and SmaI. Lane L: lambda/HindIII digest, lane 2: pEG1-06, lane 3: pEG1-06 digested with EcoRI, lane 4: pEG1-06 digested with SmaI.

Table S8. *In silico* digest predictions and observations of pCCRT11-6 digested with EcoRI

Fragment number	Fragment Size (bp)	Seen on gel image?
F8	51,264	Yes
F3	43,712	Yes
F2	11,669	Yes
F1	5,473	Yes
F4	4,628	Yes
F7	3,839	Yes
F5	600	No
F6	284	No

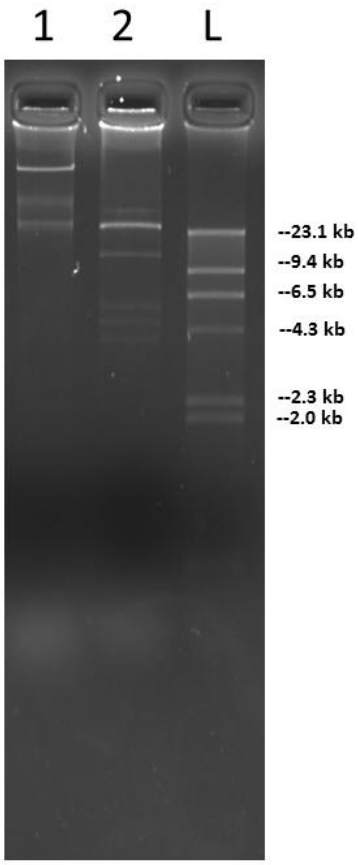

Figure S16. Restriction digestion of pCCRT11-6 with EcoRI. Lane 1: pCCRT11-6, lane 2: pCCRT11-6 digested with EcoRI, lane L: lambda/HindIII digest.

Table S9. *In silico* digest predictions and observations of pCCRT11-6 digested with HindIII

Fragment number	Fragment Size (bp)	Seen on gel image?
F1	121,469	Yes

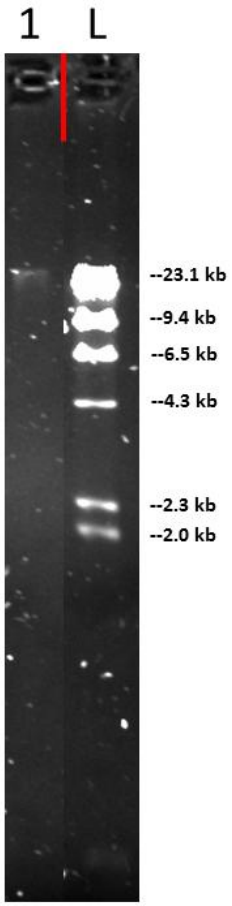

Figure S17. Restriction digestion of pCCRT11-6 with HindIII. Lane 1: pCCRT11-6 digested with HindIII, lane L: lambda/HindIII digest. Red line indicates where the image was cut to remove unnecessary lanes.

Table S10. *In silico* digest predictions and observations of pCCRT11-6 digested with SmaI

Fragment number	Fragment Size (bp)	Seen on gel image?
F3	104,834	Yes
F1	15,869	Yes
F2	766	No

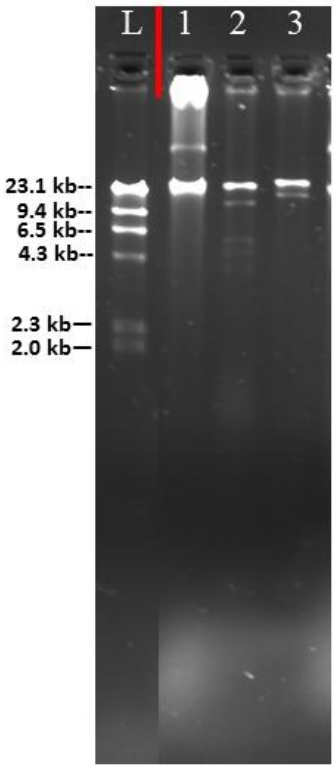

Figure S18. Restriction digestion of pCCRT11-6 with EcoRI and SmaI. Lane L: lambda/HindIII, lane 1: pCCRT11-6, lane 2: pCCRT11-6 digested with EcoRI, lane 3: pCCRT11-6 digested with SmaI. Red line indicates where the image was cut to remove unnecessary lanes.

Table S11. *In silico* digest predictions and observations of pCCP1 digested with HindIII

Fragment number	Fragment Size (bp)	Seen on gel image?
F6	22,980	Yes
F3	13,827	Yes
F4	11,659	Yes
F1	7,528	Yes
F2	2,561	Yes
F5	1,287	Yes

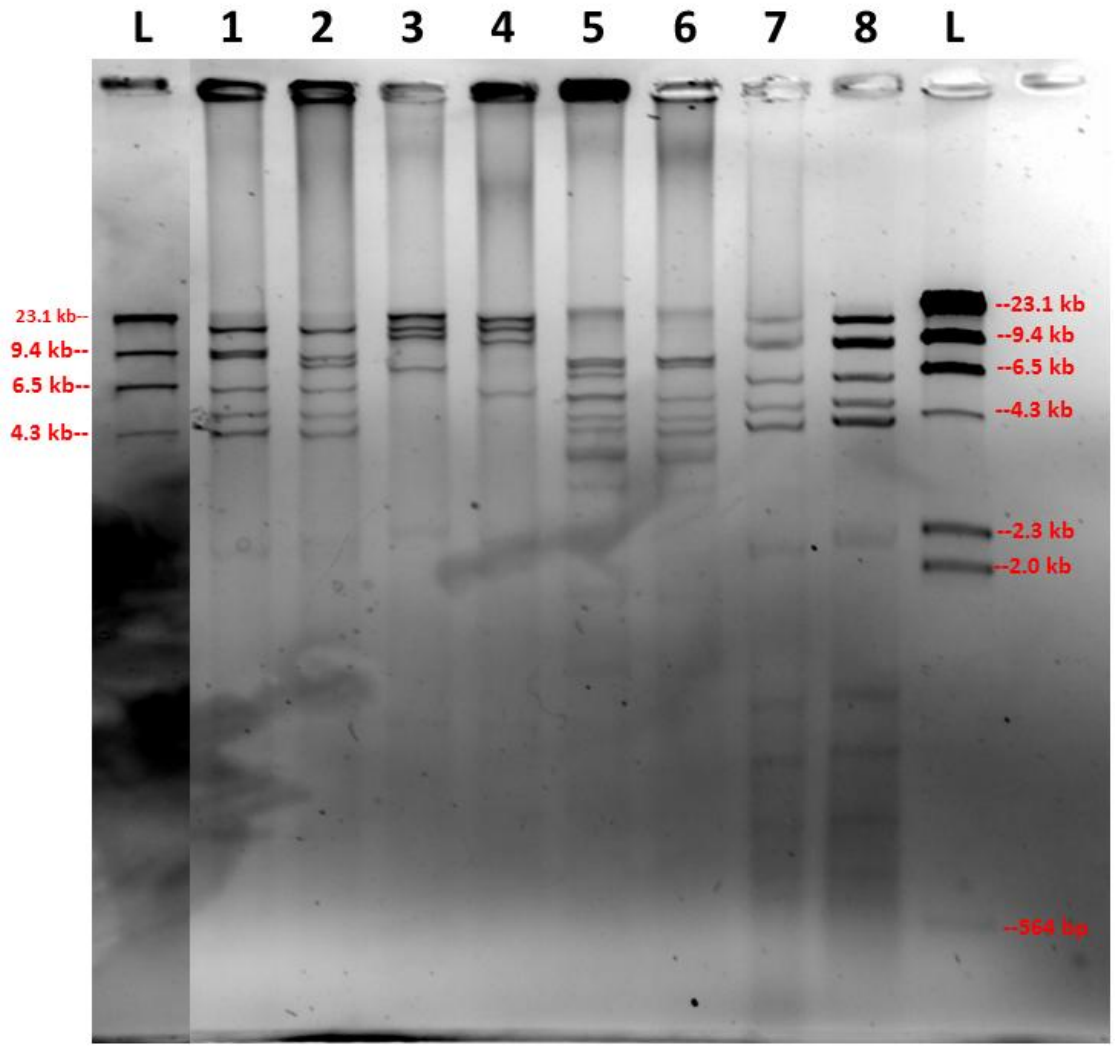

Figure S19. Restriction digestion of pCCP1 and pCCP2 with EcoRI, HindIII and SmaI. Lanes L: lambda/HindIII digest, lane 1: pCCP1 digested with EcoRI, lane 2: pCCP2 digested with EcoRI, lane 3: pCCP1 digested with HindIII, lane 4: pCCP2 digested with HindIII, lane 5: pCCP1 digested with SmaI, lane 6: pCCP2 digested with SmaI, lane 7: pCCP1 digested with EcoRI and PS treated, lane 8: pCCP1 digested with EcoRI (PS treated and Agencourt cleaned). Gel image was cut between the first ladder lane and lane 1 to remove unnecessary lanes.

Table S12. *In silico* digest predictions and observations of pCCP1 digested with SmaI

Fragment number	Fragment Size (bp)	Seen on gel image?
F3	7,963	Yes, bands overlap
F15	7,490	Yes, bands overlap
F10	6,718	Yes
F9	5,525	Yes, bands overlap

F2	5,406	Yes, bands overlap
F14	4,591	Yes
F16	4,166	Yes
F5	3,631	Yes, bands overlap
F11	3,505	Yes, bands overlap
F7	3,458	Yes, bands overlap
F13	2,997	Yes
F12	1,989	Yes
F1	1,524	Yes
F8	766	No
F6	70	No
F4	40	No

Table S13. *In silico* digest predictions and observations of pCCP2 digested with EcoRI

Fragment number	Fragment Size (bp)	Seen on gel image?
F8	14,139	Yes
F3	9,056	Yes
F5	8,142	Yes
F2	6,256	Yes
F4	4,952	Yes
F11	4,293	Yes, bands overlap
F10	4,240	Yes, bands overlap
F7	2,523	Yes
F9	2,378	Yes
F14	1,344	No
F6	845	No
F12	686	No
F13	619	No
F1	369	No

Table S14. *In silico* digest predictions and observations of pCCP2 digested with HindIII

Fragment number	Fragment Size (bp)	Seen on gel image?
F6	22,980	Yes
F3	15,261	Yes
F4	11,659	Yes
F1	6,094	Yes
F2	2,561	Yes
F5	1,287	Yes

Table S15. *In silico* digest predictions and observations of pCCP2 digested with SmaI

Fragment number	Fragment Size (bp)	Seen on gel image?
F10	8,152	Yes, bands overlap
F3	7,963	Yes, bands overlap
F15	7,490	Yes
F2	5,406	Yes
F14	4,591	Yes
F16	4,166	Yes, bands overlap (with F9)
F9	4,091	Yes, bands overlap (with F16)
F5	3,631	Yes, bands overlap (with F11 and F7)
F11	3,505	Yes, bands overlap (with F5 and F7)
F7	3,458	Yes, bands overlap (with F5 and F11)
F13	2,997	Yes
F12	1,989	Yes
F1	1,524	Yes
F8	766	No
F6	70	No
F4	43	No

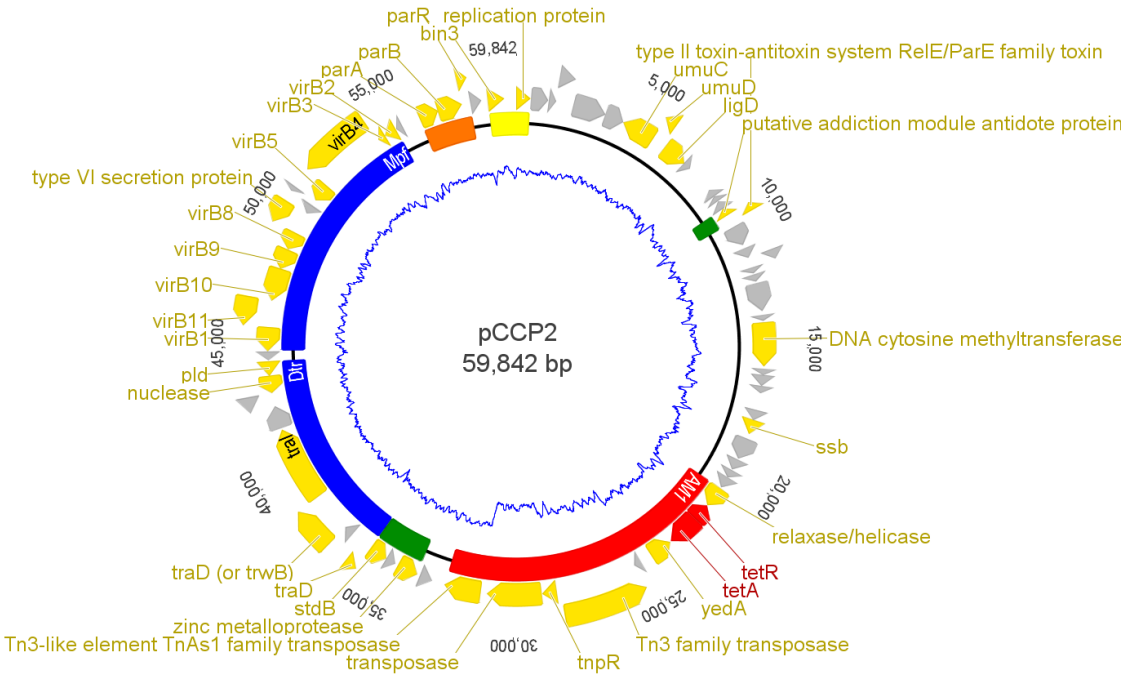

Figure S20. Functional modules and CDS mapped to the pCCP1 genome assembly. Coding sequences and annotations are on the outer circle (outermost CDS are on complementary strand and the inner CDS are on the template strand): CDS annotated by Prokka and confirmed by BLASTp search (gold), antibiotic resistance genes annotated with ABRicate (red), and CDS annotated by Prokka as “hypothetical protein” (grey). Plasmid functional modules labeled on the inner circle: Accessory Module (AM1) (red), stability and partitioning regions (orange), DNA transfer (Dtr) and mate pair formation (Mpf) regions (blue), and replication regions (yellow). Inner blue trace represents % GC. pCCP2 has a 57.09 % mean GC content.

References

Acar JF, Goldstein FW. 1996. Disk susceptibility test. In V. Lorian, (Ed.). *Antibiotics in Laboratory Medicine* pp. 1–51. Baltimore: Williams & Wilkins.

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. 2005. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl and Environ Microbiol* **71**: 8966-8969.

Adamczyk M, and Jagura-Burdzy G. 2003. Spread and survival of promiscuous IncP-1 plasmids. *Acta Biochim Pol* **50**:425-453.

Alvarado A, Garcillán-Barcia MP, de la Cruz F. 2012. A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. *PLoS One* **7:**e40438.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed Dec 3, 2017.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403–10.

Arredondo-Alonso S, Willems RJ, Schaik W, Schurch AC. 2017. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* **3**: doi: 10.1099/mgen.0.000128.

Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, Wang J, Wu M, Liu X, Tian G, Wang J, Wang J, Yang H, Zhang X. 2011. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* **12**:R95.

Botts RT, Apffel BA, Walters CJ, Davidson KE, Echols RS, Geiger MR, Guzman VL, Haase VS, Montana MA, Brown CJ, Top EM, Cummings DE. 2017. Characterization of four multidrug resistance plasmids captured from the sediments of an urban coastal wetland. *Front Microbiol* **8**:1-14.

Broom JE, Hill DF, Hughes G, Jones WA, McNaughton JC, Stockwell PA, Petersen GB. 1995. Sequence of a transposon identified as Tn1000 (γδ). *DNA Sequence* **5**:185-189.

Cantas L, Shah SQA, Cavaco LM, Manaia CM, Walsh F, Popowska M, Garelick H, Bürgmann H, Sørum H. 2013. A brief multi-disciplinary review on antimicrobial resistance in medicine and its linkage to the global environmental microbiota. *Front in Microbiol* **4:**96.

Caratolli A, Villa L, Poirel L, Bonnin RA, Nordmann P. 2012. Evolution of IncA/C blaCMY-2-carrying plasmids by acquisition of the blaNDM-1 carbapenemase gene. *Antimicrob Agents Chemother* **56**:783-786.

Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. *In silico* detection and typing of plasmids using

PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents and Chemother* **58**:3895–3903.

Chen I, Dubnau D. 2004. DNA uptake during bacterial transformation. *Nat Rev Microbiol* **2:**241-249.

Center for Veterinary Medicine, US Food and Drug Administration. 2016 Summary report on antimicrobials sold or distributed for use in food-producing animals. Accessed March 3, 2018. https://www.fda.gov/downloads/forindustry/userfees/animaldruguserfeeactadufa/ucm588085.pdf

Cury J, Jové T, Touchon M, Néron B, Rocha EPC. 2016. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* **44**:4539-4550.

Davies J, Davies D. 2010. Origins and Evolution of Antibiotic Resistance. *Microbiol and Mol Biol Reviews* **74**:417-433.

De Coster W, D'Hert S, Schultz DT, Cruts M, Ban Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* (in press, available online) https://doi.org/10.1093/bioinformatics/bty149

Dennis JJ. 2005. The evolution of IncP catabolic plasmids. *Curr Opin Biotechnol* **16**:291-298.

Desai A, Marwah V, Yadav A, Jha V, Dhaygude K, Bangar U, Kulkarni V, Jere A. 2013. Identification of optimum sequencing depth especially for *de novo* genome assembly of small genomes using next generation sequencing data. *PLoS One* **8**:e60204.

Desai KK, Miller BG. 2010. Recruitment of genes and enzymes conferring reistance to the non-natural toxin bromoacetate. *Proc Natl Acad Sci USA* **107**:17968-17973.

Doyle MP, Loneragan GH, Scott HM, Singer RS. 2013. Antimicrobial Resistance: Challenges and Perspectives. *Comp Rev in Food Sci and Food Safety* **12**:234-248.

Earl D, Bradnam K, St. John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung W, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol İ, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, Yang S, Wu W, Chou W, Srivastava A, Shaw TI, Ruby JG, Skewes-Cox P, Betegon M, Dimon MT, Solovyev V, Seledtsov I, Kosarev P, Vorobyev D, Ramirez-Gonzalez R, Leggett R, MacLean D, Xia F, Luo R, Li Z, Xie Y, Liu B, Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Yin S, Sharpe T, Hall G, Kersey PJ, Durbin R, Jackman SD, Chapman JA, Huang X, DeRisi JL, Caccamo M, Li Y, Jaffe DB, Green RE, Haussler D, Korf I, Paten B. 2011. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res* **21:**2224-2241.

Elliot I, Ming D, Robinson MT, Nawtaisong P, Newton PN, de Cesare M, Bowden R, Batty EM. 2018. MinION sequencing enables rapid whole genome assembly of Rickettsia typhi in a resource-limited setting. *bioRxiv* doi.org/10.1101/292102

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**:3047-3047.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using Phred II error probabilities. *Genome Research* **8**:186-194.

Fricke WF, Welch TJ, McDermott PF, Mammel MK, LeClerc JE, White DG, Cebula TA, Ravel J. 2009. Comparative genomics of the IncA/C multidrug resistance plasmid family. *J Bacteriol* **191**:4750-4757.

Fry JC, Day MJ. 1990. Plasmid transfer in the epilithon, p. 55-80. *In* J. C. Fry and M. J. Day (eds.), Bacterial Genetics in Natural Environments. Springer Netherlands, Dordrecht.

Gallagher, B. 2007. Detecting and quantifying enterohemorrhagic *E. coli* and *Salmonella* genes in Virginia streams. M.S. Thesis, James Madison University, Harrisonburg, VA

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**:201-206.

Garcillan-Barcia MP, Ruiz del Castillo B, Alvarado A, de la Cruz F, Martnez-Martnez L. 2015. Degenerate primer MOB typing of multiresistant clinical isolates of *E. coli* uncovers new plasmid backbones. *Plasmid* **77**:17-27.

Gehr E. 2013. The potential for replication and transmission of antibiotic resistance plasmids in an *E. coli* population in agriculturally impacted stream sediment, M.S. Thesis. James Madison University, Harrisonburg, VA.

George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N, Sheppard AE, Mathers A, Norris A, Navickaite I, Eaton C, Iqbal Z, Crook DW, Phan HTT. 2017. Resolving plasmid structures in *Enterobacteriaceae* using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. *Microb Genom* **3**:1-8.

Gotz A, Smalla K. 1997. Manure enhances plasmid mobilization and survival of *Pseudomonas putida* introduced into field soil. *Appl and Environ Microbiol* **63**:1980-1986

Greated A, Lamberten L, Williams PA, Thomas CM. 2002. Complete sequence of the IncP-9 TOL plasmid pWW0 from *Pseudomonas putida*. *Envir Microbiol* **4**:856-871.

Harayama S, Lehrbach PR, Timmis KN. 1984. Transposon mutagenesis analysis of meta-cleavage pathway operon genes of the TOL plasmid of *Pseudomonas putida* mt-2. *J Bacteriol* **160**:251-255.

Harmer CJ, Hall RM. 2015. The A to Z of A/C plasmids. *Plasmid* **80**:63-82.

Herrick JB, Haynes R, Heringa S, Brooks JM, Sobota LT. 2014. Coselection for resistance to multiple late-generation human therapeutic antibiotics encoded on tetracycline resistance plasmids captured from uncultivated stream and soil bacteria. *J Appl Microbiol* **117**:380-389.

Herrick JB, Turner SD, Gehr E, Libuit K, Kapsak CJ. 2015. Nanopore sequencing of transmissible tetracycline plasmids captured without cultivation from stream sediment reveals linked genes encoding resistance to multiple human clinical antibiotics. Poster presented at the ASM Conference "Rapid Next-Generation Sequencing and Bioinformatic Pipelines for Enhanced Molecular Epidemiologic Investigation of Pathogens", Sept. 2015. doi:10.6084/m9.figshare.1564751.

Illumina. 2016. Optimizing Cluster Density on Illumina Sequencing Systems. Pub. No. 770-2014-038

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore Sequencing and assembly of a human genome with ultra-long reads. *Nat Methods* **36**:338-345.

Jain M, Olsen HG, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**:239.

Kamada K, Hanaoka F, Burley SK. 2003. Crystal structure of the mazE/mazF complex: molecular bases of antidote-toxin recognition. *Mol Cell* **11**:875:884.

Kav AB, Benhar I, Mizrahi I. 2013. A method for purifying high quality and high yield plasmids DNA for metagenomic and deep sequencing approaches. *J of Microbiol Methods* **95**:272-279.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649.

Leggett RM, Clark MD. 2017. A world of opportunities with nanopore sequencing. *J Exp Bot* **68**:5419-5429.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078-2079.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997v2 [q-bio.GN].

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* (in press, available online) https://doi.org/10.1093/bioinformatics/bty191

Li J, Nation RL, Turnidge JD, Milne RW, Coulthard K, Rayner CR, Paterson DL. 2006. Colistin: the re-emerging antibiotic for multidrug-resistant Gram-negative bacterial infections. *Lancet Infect Dis* **6**:589-601.

Li R, Xie M, Dong N, Lin D, Yang X, Wong MHY, Chan EWC, Chen S. Efficient generation of complete sequences of MDR-encoding plamids by rapid assembly of MinION barcoding sequencing data. *GigaScience* **7**:1-9.

Libuit K. 2016. Next-generation sequencing of a multi-drug resistance plasmid captured from stream sediment. M.S. Thesis. James Madison University, Harrisonburg, VA.

Liu Y, Feng Y, Zhang X, McNally A, Zong Z. 2017. A new variant mcr-3 in an extensively drug-resistant *Escherichia coli* clinical isolate carrying mcr-1 and blaNDM-5. *Antimicrob Agents Chemother* **61**:e01757-17.

Liu Y, Wang Y, Walsh TR, Yi L, Zhang R, Spencer J, Doi Y, Tian G, Dong B, Huang X, Yu L, Gu D, Ren H, Chen X, Lv L, He D, Zhou H, Liang Z, Liu J, Shen J. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect Dis* **16**:161-168.

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled *de novo* using only nanopore sequence data. *Nat Methods* **12**:733-735.

Lorenz MG, Wackernagel W. 1994. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev* **58**:563-602.

Lu X, Hu Y, Luo M, Zhou H, Wang X, Du Y, Li Z, Xu J, Zhu X, Kan B. 2016. MCR-1.6, a new MCR variant carried by an IncP plamsmid in a colistin-resistant *Salmonella enterica* serovar *Typhimurium* isolate from a healthy individual. *Antimicro Agents Chemother* **61**:e02632-16.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, RonanMT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flowcell. *Nat Commun* **9**:541.

Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A. 2016. Icarus: visualizer for de novo assembly evaluation, *Bioinformatics* **32**: 3321-3323.

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinformatics* **14**:193-202.

Moura A, Soares M, Pereira C, Leitão N, Henriques I, Correia A. 2009. INTEGRALL: a database and search engine for integrons, integrases, and gene cassettes. *Bioinformatics* **25**:1096-1098.

Norman A, Riber L, Luo W, Li LL, Hansen LH, Sorensen SJ. 2014. An improved method for including upper size range plasmid in metamobiliomes. *PLoS One* **9**:e104405.

Novick RP, Clowes RC, Cohen SN, Curtiss R, Datta N, Falkow S. 1976. Uniform nomenclature for bacterial plasmids: a proposal. *Bacteriol Rev* **40**:168–89.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405:**299-304.

Oxford Nanopore Technologies 2016. In brief: Oxford Nanopore updates at Nanopore Community meeting in New York. https://nanoporetech.com/about-us/news/brief-oxford-nanopore-updates-nanopore-community-meeting-new-york Accessed June 19th, 2018.

Quick J, Loman NJ, *et al*. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**:228.

Loman NJ, Pallen MJ. 2015. Real-time, portable genome sequencing for Ebola sequencing. *Nat Rev Microbiol* **13**:787.

Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore Sequencing and assembly of a human genome with ultra- long reads. *Nat Methods* **36**:338-345.

Rademaker JLW, Louws FJ, Bruijn FJD. 1998. Characterization of the diversity of ecologically important microbes by rep-PCR fingerprinting. *In*: Molecular Microbial Ecology Manual. p. 1–26. (Akkermans, A.D.L., Elsas, J.D.v., & Bruijn, F.J.d., Editors). Dordrecht, Netherlands: Kluwer Academic Publishers.

Revilla C, Garcillán-Barcia MP, Fernandez-Lopez R, Thomson NR, Sanders M, Cheung M, Thomas CM, De La Cruz F. 2008. Different pathways to acquiring resistance genes illustrated by the recent evolution of IncW plasmids. *Antimicrob Agents Chemother* **52**:1472-1480.

Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, Guerra B, Mevius DJ, Hordijk J. 2018. Plasmids carrying antimibrobial reistance genes in Enterobacteriaceae. *J Antimicrob Chemother* **73**:1121-1137.

San Milan A. 2018. Evolution of plasmid-mediated antibiotic resistance in the clinical context. *Trends Microbiol* in press. 10.1016/j.tim.2018.06.007

Sambrook J. 2006. The condensed protocols from molecular cloning: a laboratory manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2006.

Schlüter A, Szczepanowski R, Pühler A, Top EM. 2007. Genomics of IncP-1 antibiotic resistance plasmids isolated from wastewater treatment plants provides evidence for a widely accessible drug resistance gene pool. *FEMS Microbiol Rev* **31**:449-477.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068-2069.

Sevastsyanovich YR, Titok MA, Krasowiak R, Bingle LEH, Thomas CM. 2005. Ability of IncP-9 plasmid pM3 to replicate in *Escherichia coli* is dependent on both *rep* and *par* functions. *Mol Microbiol* **57**:819-833.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26**:1135-1145.

Shoemaker NB, Vlamakis KH, Salyers AA. 2001. Evidence for extensive resistance gene transfer among Bacteroides spp. and among *Bacteroides* and other genera in the human colon. *Appl and Environ Microbiol* **67**:561-568.

Smalla K, Jechalke S, Top EM. 2015. Plasmid Detection, Characterization, and Ecology, p 445-458. *In* Tolmasky M, Alonso J (ed), *Plasmids: Biology and Impact in Biotechnology and Discovery.* ASM Press, Washington, DC.

Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. 2010. Mobility of plasmids. *Microbiol Mol Biol Rev* **74**:434-452.

Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**:1009-1010.

Szabó M, Nagy T, Wilk T, Farkas T, Hegyi A, Olasz F, Kiss J. 2016. Characterization of two multidrug-resistant IncA/C plasmids from the 1960s by using the MinION sequencer device. *Antimicrob Agents Chemother* **60**:6780–6786.

Thorsted PB, Macartney DP, Akhtar P, Haines AS, Ali N, Davidson P, Stafford T, Pocklington MJ, Pansegrau W, Wilkins BM, Lanka E, Thomas CM. 1998. Complete sequence of IncPβ plasmid R751: implications for evolution and organization of the IncP backbone. *J Mol Biol* **282**:969-990.

Turcatti G, Romieu A, Fedurco M, Tairi A. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res* **34**:e25.

Virginia Department of Environmental Quality. 2016. Virginia water quality assessment 305(b)/303(d) integrated report. Available at: http://www.deq.virginia.gov/Portals/0/DEQ/Water/WaterQualityAssessments/IntegratedReport/2016/ir16_Integrated_Report_Full_Draft.pdf. Accessed March 5th, 2018

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for

comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**:e112963.

Wang G, Xu X, Chen J, Berg DE, Berg CM. 1994. Inversions and deletions generated by a mini- γδ (Tn1000) transposon. *J Bacteriol* **176**:1332-1338.

Wegener HC. 2003. Antibiotics in animal feed and their role in resistance development. *Curr Opin in Microbiol* **6**:439-445.

Wick LM, Qi W, Lacher DW, Whitman TS. 2005. Evolution of genomic content in the stepwise emergence of *Eschericha coli* O157:H7. *J Bacteriol* **187**:1783-1791.

Wick RR Judd LM, Holt KE. 2018. Comparison of Oxford Nanopore basecalling tools. (Version v5.1). Zenodo. http://doi.org/10.5281/zenodo.1188469

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* **3**:1-7.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PloS Comput Biol* **13**:e1005595.

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**:3350-3352.

Williams LE, Detter C, Barry K, Lapidus A, Summers AO. 2006. Facile recovery of individual high-molecular-weight, low-copy-number natural plasmids for genomic sequencing. *Appl Environ Microbiol* **72**:4899-4906.

Yerushalmi H, Lebendiker M, Schuldiner S. 1995. EmrE, an *Escherichia coli* 12-kDa multidrug transporter, exchanges toxic cations and H^+ and is soluble in organic solvents. *J Biol Chem* **270**:6856-6863.